

NETWORK GRANGER CAUSALITY WITH INHERENT GROUPING STRUCTURE

BY SUMANTA BASU^{*}, ALI SHOJAIE[†] AND GEORGE MICHAILIDIS^{*}

University of Michigan^{} and University of Washington[†]*

The problem of estimating high-dimensional network models arises naturally in the analysis of many physical, biological and socio-economic systems. Examples include stock price fluctuations in financial markets and gene regulatory networks representing effects of regulators (transcription factors) on regulated genes in genetics. We aim to learn the structure of the network over time employing the framework of Granger causal models under the assumptions of sparsity of its edges and inherent grouping structure among its nodes. We introduce a thresholded variant of the Group Lasso estimator for discovering Granger causal interactions among the nodes of the network. Asymptotic results on the consistency of the new estimation procedure are developed. The performance of the proposed methodology is assessed through an extensive set of simulation studies and comparisons with existing techniques.

1. Introduction. Granger causality [Granger, 1969] provides a statistical framework for determining whether a time series X is useful in forecasting another one Y , through a series of statistical tests. It has found wide applicability in economics, including testing relationships between money and income [Sims, 1972], government spending and taxes on economic output [Blanchard and Perotti, 2002], stock price and volume [Hiemstra and Jones, 1994], etc.

Extensions involving multiple time series can be handled through analysis of vector autoregressive processes (VAR) [Lütkepohl, 2005], which provide a convenient framework for analysis of relationships amongst multiple variables. As a result, the Granger causality framework has recently found diverse applications in biological sciences including genetics, bioinformatics and neurosciences to understand the structure of gene regulation, protein-protein interactions and brain circuitry, respectively. In these applications, the main goal is to reconstruct a network of interactions amongst the entities involved based on time course data.

AMS 2000 subject classifications: Primary 62J07, 62M10, 62M15

Keywords and phrases: Granger causality, high dimensional networks, vector autoregressive model, panel study, group lasso, thresholding

It should be noted that the concept of Granger causality is based on associations between times series, and only under very stringent conditions, true causal relationships can be inferred [Pearl, 2000]. Nonetheless, this framework provides a powerful tool for understanding the interactions among random variables based on time course data.

Network Granger causality (NGC) extends the notion of Granger causality among two variables to a wider class of p variables. More generally if X_1^t, \dots, X_p^t are p stationary time series, with $\mathbf{X}^t = (X_1^t, \dots, X_p^t)'$, we consider the class of models

$$(1.1) \quad \mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \dots + A^d \mathbf{X}^{T-d} + \epsilon^T,$$

where d the order of the VAR model is allowed to be unknown and the innovation process satisfies $\epsilon^T \sim N(0, \sigma^2 I)$. We call A^1, \dots, A^d the adjacency matrices from lags $1, \dots, d$. In this model, X_j^t is said to be Granger causal for X_i^T if $A_{i,j}^t$ is statistically significant. In this case, there exists an edge $X_j^t \rightarrow X_i^T$ in the underlying network model comprising of $T \times p$ nodes (see Figure 1). Note that the presence of ordering between the variables in this

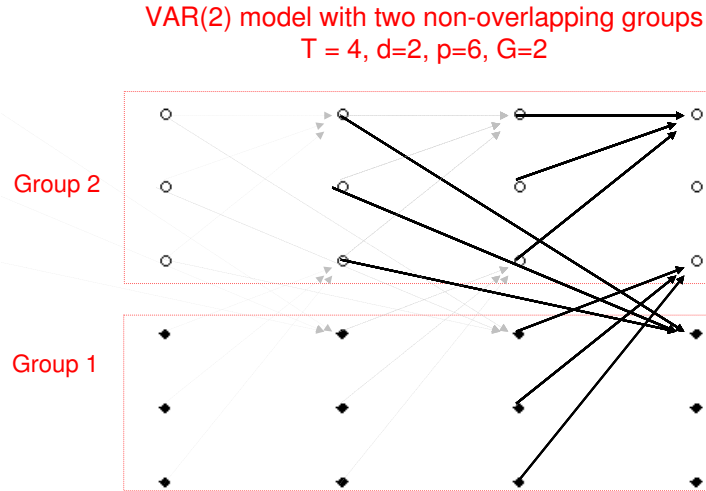


Fig 1: An Example of a network Granger model with two non-overlapping groups observed over $T = 4$ time points

network, due to their temporal structure, simplifies significantly the network estimation problem [Shojaie and Michailidis, 2010a]. Nevertheless, one still has to deal with estimating a high-dimensional network (e.g. hundreds of genes) from a limited number of samples.

Estimation of NGC models often arises in the analysis of large panel data in econometrics, where one is interested to understand the temporal relationship of several economic variables observed over time across a panel of subjects. Such an example is presented in Section 6.1 that examines the structure of the balance sheets of the 50 largest US banks by size, over 9 quarterly periods. The nature of high-dimensionality in this problem comes from both estimation of p^2 coefficients for the adjacency matrices A^1, \dots, A^d , but also from the fact that the order of the time series d is often unknown. Thus, in practice, one must either “guess” the order of the time series (often times, it is assumed that the data is generated from a VAR(1) model, which can result in significant loss of information), or include all of the past time points, resulting in significant increase in the number of variables in cases where $d \ll T$. Thus, efficient estimation of the order of the time series becomes crucial.

Recent work of [Fujita et al. \[2007\]](#) and [Lozano et al. \[2009\]](#) employed NGC models coupled with penalized ℓ_1 regression methods to learn gene regulatory mechanisms from time course microarray data. Specifically, [Lozano et al. \[2009\]](#) proposed to group all the past observations, using a variant of group lasso penalty, in order to construct a relatively simple Granger network model. This penalty takes into account the average effect of the covariates over different time lags and connects Granger causality to this average effect being significant. However, it suffers from significant loss of information and makes the consistent estimation of the signs of the edges difficult (due to averaging). [Shojaie and Michailidis \[2010b\]](#) proposed a truncating lasso approach by introducing a truncation factor in the penalty term, which strongly penalizes the edges from a particular time lag, if it corresponds to a highly sparse adjacency matrix.

Despite recent use of NGC in high dimensional settings, theoretical properties of the resulting estimators have not been fully investigated. For example, [Lozano et al. \[2009\]](#) and [Shojaie and Michailidis \[2010b\]](#) discuss consistency of the resulting estimators, but neither address in depth selection consistency properties nor do they examine under what vector autoregressive structures the obtained results hold. Hence, there is significant room for theoretical work in understanding theoretically the performance of penalized estimators in NGC models.

In addition, in many applications structural information about the variables exists, which could improve the estimation of Granger causal models. For example, genes can be naturally grouped according to their function or chromosomal location, stocks according to their industry sectors, assets/liabilities according to their class, etc. This information can be incor-

porated to the Granger causality framework through a group lasso penalty. If the group specification is correct it enables estimation of denser networks with limited sample sizes [Bach, 2008, Huang and Zhang, 2010, Lounici et al., 2011]. However, the group lasso penalty can achieve model selection consistency only at a group level. In other words, if the groups are misspecified, this procedure can not perform within group variable selection [Huang et al., 2009], an important feature in many applications. To address this issue, we propose a new notion of “direction consistency”, and use this notion to introduce a thresholded variant of group lasso for NGC models.

In this paper, we develop a general framework that accommodates different variants of group lasso penalties for NGC models. It allows for the simultaneous estimation of the order of the times series and the Granger causal effects; further, it allows for variable selection even when the groups are misspecified. In summary, the key contributions of this work are: (i) investigate sufficient conditions that explicitly take into consideration the structure of the VAR(d) model to establish norm and variable selection consistency, (ii) introduce the novel notion of direction consistency, which generalizes the concept of sign consistency, and use it to establish variable selection consistency of group lasso estimates with misspecified group structures, and (iii) use the latter notion to introduce an easy to compute thresholded variant of group lasso, that performs within group variable selection in addition to group sparsity pattern selection. Application of the proposed framework to data from banks’ balance sheets and temporal regulatory mechanisms related to T-cell activation indicates that the resulting estimates provide novel insight into interactions among components of the system, as well as improved prediction of future values of the variables.

The rest of the paper is organized as follows. In Section 2, we formulate the group NGC estimate and its variants. We explain their major advantages and briefly discuss the implementation procedure. Section 3 describes the notation used and introduces the notion of direction consistency, and discusses different assumptions required for the consistency of NGC estimates. The theoretical properties of group NGC estimates are discussed in Section 4, where non-asymptotic bounds for their norm and variable selection consistency are established. Section 5 reports the results of numerical experiments, under different settings, and Section 6 applies the different NGC methods on two real datasets.

2. Model and Framework.

2.1. *Notation.* Consider a VAR model

$$(2.1) \quad \underbrace{\mathbf{X}^T}_{p \times 1} = \underbrace{A^1}_{p \times p} \mathbf{X}^{T-1} + \dots + A^d \mathbf{X}^{T-d} + \epsilon^T$$

observed over T time points $t = 1, \dots, T$, with innovation process $\epsilon^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{p \times p})$. The index set of the variables $\mathbb{N}_p = \{1, 2, \dots, p\}$ can be partitioned into G non-overlapping groups \mathcal{G}_g , i.e., $\mathbb{N}_p = \cup_{g=1}^G \mathcal{G}_g$ and $\mathcal{G}_g \cap \mathcal{G}_{g'} = \emptyset$ if $g \neq g'$ and where $k_g = |\mathcal{G}_g|$ denotes the size of the g^{th} group with $k_{max} = \max_{1 \leq g \leq G} k_g$.

For any matrix A , we denote the i^{th} row by $A_{i:}$, j^{th} column by $A_{:,j}$ and the collection of rows (columns) corresponding to the g^{th} group by $A_{[g]:}$ ($A_{:, [g]}$). The transpose of a matrix A is denoted by A' and its Frobenius norm by $\|A\|_F$. The symbol $A^{1:T}$ is used to denote the concatenated matrix $[A^1 : \dots : A^T]$. Further, for notational convenience, we reserve the symbol $\|\cdot\|$ to denote the ℓ_2 norm of a vector and/or the spectral norm of a matrix. Any other norm will be indexed explicitly (e.g., $\|\cdot\|_1$, $\|\cdot\|_{2,2}$, $\|\cdot\|_{2,\infty}$) to avoid confusion. Also for any vector β , we use β_j to denote its j^{th} coordinate and $\beta_{[g]}$ to denote the coordinates corresponding to the g^{th} group.

2.2. *Network Granger causal (NGC) estimates with group sparsity.* Consider n replicates from the NGC model (2.1), and denote the $n \times p$ observation matrix at time t by \mathcal{X}^t . For example in a panel-VAR setting, the data on p economic variables on n subjects (firms, households etc.) can be observed over T time points. The data is high-dimensional if either T or p is large compared to n . In such a scenario, we assume the existence of an underlying group sparse structure, i.e., the support of each row of $A^{1:T} = [A^1 : \dots : A^T]$ in the model (2.1) can be covered by a small number of groups s , where $s \ll (T-1)G$. Note that the groups can be misspecified in the sense that the coordinates of a group covering the support need not be all non-zero. Hence, for a properly specified group structure we shall expect $s \ll \|A_{i:}^{1:T}\|_0$. On the contrary, with many misspecified groups, s can be of the same order, or even larger than $\|A_{i:}^{1:T}\|_0$.

The group Granger causal estimates of the adjacency matrices A^1, \dots, A^{T-1} are obtained by solving the following optimization problem

$$(2.2) \quad \hat{A}^{1:T-1} = \underset{A^1, A^2, \dots, A^{T-1} \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} \frac{1}{2n} \left\| \mathcal{X}^T - \sum_{t=1}^{T-1} \mathcal{X}^{T-t} (A^t)' \right\|_F^2 \\ + \lambda_n \sum_{t=1}^{T-1} \Psi^t \sum_{i=1}^p \sum_{g=1}^G w_{i,g}^t \|A_{i:[g]}^t\|_2,$$

where \mathcal{X}^t is the $n \times p$ observation matrix at time t , constructed by stacking n i.i.d. replicates from the model (2.1), w^t is a $p \times G$ matrix of suitably chosen weights, and Ψ^t is a truncating or thresholding factor, for every t . This optimization problem can be separated into the following p different penalized regression problems - for $i = 1, \dots, p$,

$$(2.3) \quad \hat{A}_{i:}^{1:T-1} = \underset{\theta^1, \theta^2, \dots, \theta^{T-1} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\mathcal{X}_{i:}^T - \sum_{t=1}^{T-1} \mathcal{X}^{T-t} \theta^t\|_2^2 \\ + \lambda_n \sum_{t=1}^{T-1} \Psi^t \sum_{g=1}^G w_{i,g}^t \|A_{i:[g]}^t\|_2.$$

The order d of the VAR model is estimated as $\hat{d} = \max_{1 \leq t \leq T-1} \{t : \hat{A}^t \neq \mathbf{0}\}$.

Different choices of weights $w_{i,g}^t$ and truncating/thresholdings factor Ψ^t introduce different variants of NGC estimates:

1. **Regular:** The regular NGC estimates correspond to the choices $\Psi^t = 1$, $w_{i,g}^t = 1$ or $\sqrt{k_g}$. The estimation procedure requires solving p group lasso penalized regression problems, as described in Section 3. Estimation and selection properties of the estimates are discussed in Section 4.1 under different choices of tuning parameter λ_n and weights w^t . In practice, λ_n can be tuned through cross-validation, that showed promising results in our numerical work.
2. **Adaptive:** The adaptive version of NGC estimates corresponds to the choices $w_{i,g}^t = \min\{1, \|\tilde{A}_{i:[g]}^t\|_2^{-1}\}$, where \tilde{A}^t are the estimates from Regular NGC. This variant of NGC involves a two-stage estimation procedure. In the first stage, only estimates of the adjacency matrices A^t are obtained, but not of the order d . The second stage uses the first-stage estimates to select weights $w_{i,g}^t$ and yields an improved rate of false positives. The algorithm requires solving p adaptive group lasso problems. The adaptive NGC estimation procedure requires a single tuning parameter λ_n which is selected in the same way as in regular NGC. Consistency of adaptive group NGC estimates rely on the consistency of adaptive group lasso estimates [cf. Wei and Huang, 2010].
3. **Thresholded:** Thresholded NGC estimates are also calculated by a two-stage procedure. The first stage involves a regular NGC estimation procedure, while at the second stage, bi-level thresholding is used. At first, the estimated groups with ℓ_2 norm less than a threshold ($\delta_{grp} = t\lambda$, $t > 0$) are set to zero. The second thresholding (within groups) is applied if the *a priori* available grouping information is not

reliable. The members within each estimated parent group are thresholded using $\delta_{misspec} = \delta_n$ for some $\delta_n \in (0, 1)$. Mathematically, for every $t = 1, \dots, T-1$, if $j \in \mathcal{G}_g$,

$$\hat{A}_{ij}^t = \tilde{A}_{ij}^t I \left\{ \left| \tilde{A}_{ij}^t \right| \geq \delta_{misspec} \left\| \tilde{A}_{i:[g]}^t \right\|_2 \right\} I \left\{ \left\| \tilde{A}_{i:[g]}^t \right\|_2 \geq \delta_{grp} \right\}$$

4. **Truncating:** A truncating variant of NGC estimates encourages accurate order selection in NGC problems, if the Granger causal effects decay over time. Truncating NGC estimates are obtained by solving a non-convex optimization problem via an iterative procedure based on a Block-Relaxation algorithms suggested in [Shojaie and Michailidis \[2010b\]](#). This variant corresponds to the choices

$$\Psi^1 = 1, \Psi^t = \exp [\Delta n I \{ \sum_{g=1}^G I_{\{\|A_{i:[g]}^{t-1}\|_0 > 0\}} < G^2 \beta / (T-t) \}], t \geq 2.$$

Consistent estimation and selection properties of truncating NGC estimates (without any group structure) were discussed in [\[Shojaie and Michailidis, 2010b\]](#) under a decay assumption on the Granger causal effects. Similar properties can be established using the consistency of regular group NGC estimates discussed in Section 4, but are not pursued in this paper.

3. Assumptions and Conditions. Note that to obtain the solution of the NGC problem, one needs to solve for each $i = 1, \dots, p$ a generic group lasso problem of the form

$$\begin{aligned} \mathbf{Y}_{n \times 1} &= \mathbf{X}_{n \times \bar{p}} \beta_{\bar{p} \times 1}^0 + \epsilon, & \epsilon &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n}) \\ \{1, \dots, \bar{p}\} &= \cup_{g=1}^{\bar{G}} \mathcal{G}_g, & |\mathcal{G}_g| &= k_g \\ (3.1) \quad \hat{\beta} &= \underset{\beta \in \mathbb{R}^{\bar{p}}}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X} \beta^0\|_2^2 + \sum_{g=1}^{\bar{G}} \lambda_g \|\beta_{[g]}\|_2 \end{aligned}$$

with $\mathbf{Y} = \mathcal{X}_i^T$, $\mathbf{X} = [\mathcal{X}^1 : \dots : \mathcal{X}^{T-1}]$, $\beta^0 = \operatorname{vec}(A_{i:}^{1:(T-1)})$, $\bar{p} = (T-1)p$, $\bar{G} = (T-1)G$ and $\lambda_g = \lambda_n w_{i,g}$. For ease of presentation, in the remainder we use p instead of \bar{p} and G instead of \bar{G} when examining the properties of the above problem.

Next, we introduce assumptions needed for establishing norm and variable selection consistency for estimators of (3.1). Specifically, for norm consistency, group variants of compatibility and restricted eigenvalue conditions are used, while selection consistency relies on group irrepresentable ones. Further, for the problem at hand, we establish a connection between group irrepresentable and group compatibility conditions (Appendix D).

Note that selection consistency of group lasso estimators involves both group-level, as well as within-groups selection consistency. Furthermore, due

to its inability to perform within group variable selection, group lasso estimates are not sign consistent whenever the groups are misspecified. Towards this end, the notion of “direction consistency” is introduced (Section 3.1.1) and the necessity of group (weak) irrepresentable conditions is established (Appendix D).

3.1. Direction Consistency and Irrepresentable Conditions.

3.1.1. Direction Consistency. As discussed in the introductory section, lasso estimates exhibit the right sparsity pattern and corresponding signs of the support variables with high probability. However, group lasso achieves sparsity at the group level [Huang et al., 2009], but not necessarily within the group itself. Hence, within group selection consistency is still unclear and several alternative penalized regression procedures have been proposed to overcome this shortcoming [Breheny and Huang, 2009, Huang et al., 2009, Zhao, Rocha and Yu, 2009]. We formulate a generalized notion of sign consistency, henceforth referred as “direction consistency”, that provides insight into the properties of group lasso estimates within a single group. Subsequently, these properties are used in a simple thresholding variant of the group lasso estimates that achieves within group variable selection consistency.

Consider a generic group lasso estimate as in (3.1). Without loss of generality, let $S = \{1, \dots, s\}$, and denote the group indices by $\text{support}(\beta^0)$, i.e.,

$$\beta^0 = [\beta_{[1]}^0, \dots, \beta_{[s]}^0, \mathbf{0}, \dots, \mathbf{0}], \quad \beta_{[g]}^0 \neq \mathbf{0} \quad \forall g \in S = \{1, \dots, s\}, \quad \sum_{g \in S} k_g = q.$$

For a vector $\tau \in \mathbb{R}^m \setminus \{\mathbf{0}\}$ we define the following quantities: $D(\tau) = \frac{\tau}{\|\tau\|_2}$ and $D(\mathbf{0}) = \mathbf{0}$. In general, the function $D(\cdot)$ indicates the direction of the vector τ in \mathbb{R}^m . Specifically, for the problem at hand, for a group $g \in S$ of size m , $D(\beta_{[g]}^0)$ indicates the direction of influence of $\beta_{[g]}^0$ at a group level as it reflects the relative importance of the influential group members. Note that for $m = 1$ the function $D(\cdot)$ simplifies to the usual $\text{sgn}(\cdot)$ function.

We define an estimate $\hat{\beta}$ as **direction consistent** at a rate δ_n , if there exists a sequence of positive real numbers $\delta_n \rightarrow 0$ such that

$$(3.2) \quad \mathbb{P} \left(\|D(\hat{\beta}_{[g]}) - D(\beta_{[g]}^0)\|_2 < \delta_n, \quad \forall g \in S \right) \rightarrow 1 \text{ as } n, p \rightarrow \infty.$$

It readily follows from the definition, that if $\hat{\beta}$ is direction consistent and $\tilde{S}_g^n = \{j \in \mathcal{G}_g : \frac{|\hat{\beta}_j|}{\|\hat{\beta}_{[g]}\|_2} > \delta_n\}$ denotes a collection of influential group mem-

bers within a group \mathcal{G}_g , which are detectable with a sample size of n , then

$$(3.3) \quad \mathbb{P}(\text{sgn}(\hat{\beta}_j) = \text{sgn}(\beta_j), \forall j \in \tilde{S}_g^n, \forall g \in \{1, \dots, s\}) \rightarrow 1 \text{ as } n, p \rightarrow \infty.$$

REMARK 3.1. *The latter observation connects the precision of group lasso estimates to the accuracy of a priori available grouping information. In particular, if the pre-specified grouping structure is correct, i.e., all the members within a group have non-zero effect, then for a sufficiently large sample size we have $\tilde{S}_g^n = \mathcal{G}_g$ and group lasso correctly estimates the sign of all the coordinates. On the other hand, in case of a misspecified a priori grouping structure, in the form of numerous zero coordinates, β_g , group lasso correctly estimates only the signs of strongly influential group members detectable with sample size n .*

3.1.2. *Group Irrepresentable Conditions.* Irrepresentable conditions are common in the literature of high-dimensional regression problems [Zhao and Yu, 2006, van de Geer and Bühlmann, 2009] and are shown to be sufficient (and essentially necessary) for selection consistency of the lasso estimates. Further these conditions are known to be satisfied with high probability, if the population analogue of the Gram matrix belongs to the Toeplitz family. Specifically, if the predictor variables in a group lasso regression problem are generated from an AR process, the design matrix satisfies irrepresentable conditions with high probability. Since we are working with vector AR processes and the population analogue of the Gram matrix $\text{var}(\mathbf{X}^{1:T})$ is block Toeplitz, the irrepresentable assumptions are natural candidates for studying selection consistency of the estimates. Next, we formulate group analogues of these conditions.

Consider the setup of a group lasso penalized linear model in (3.1) with p regressors partitioned into G groups, of which only the first s groups (of total size q) exert non-zero signal (influence) on the response. We partition the design matrix and the coefficient vector into signal and non-signal parts

$$(3.4) \quad \underbrace{\mathbf{X}}_{n \times p} = \left[\underbrace{\mathbf{X}_{(1)}}_{n \times q} : \underbrace{\mathbf{X}_{(2)}}_{n \times (p-q)} \right]$$

$$(3.5) \quad \underbrace{\beta^0}_{p \times 1} = \left[\underbrace{\beta_{[1]}^0, \dots, \beta_{[s]}^0}_{k_1 + \dots + k_s = q}, \underbrace{\mathbf{0}, \dots, \mathbf{0}}_{p-q} \right] = [\beta_{(1)}^0 : \beta_{(2)}^0]$$

$$(3.6) \quad C = \frac{1}{n} \mathbf{X}' \mathbf{X} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

Also, for a q -dimensional vector θ define the stacked direction vectors

$$(3.7) \quad \underbrace{\tilde{D}(\tau)}_{q \times 1} = \begin{bmatrix} \underbrace{D(\tau_{[1]})}_{k_1 \times 1} \\ \vdots \\ \underbrace{D(\tau_{[s]})}_{k_s \times 1} \end{bmatrix}, \quad K = \begin{bmatrix} \lambda_1 \mathbf{I}_{k_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 \mathbf{I}_{k_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_s \mathbf{I}_{k_s} \end{bmatrix}$$

Uniform Irrepresentable Condition is satisfied if there exists $0 < \eta < 1$ such that for all $\tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \leq g \leq s} \|\tau_{[g]}\|_2 \leq 1$

$$(3.8) \quad \frac{1}{\lambda_g} \left\| \left[C_{21}(C_{11})^{-1} K \tau \right]_{[g]} \right\| < 1 - \eta, \quad \forall g \notin S = \{1, \dots, s\}$$

Weak Irrepresentable Condition is satisfied if

$$(3.9) \quad \frac{1}{\lambda_g} \left\| \left[C_{21}(C_{11})^{-1} K \tilde{D}(\beta_{(1)}^0) \right]_{[g]} \right\| \leq 1, \quad \forall g \notin S = \{1, \dots, s\}$$

Note that these definitions revert to usual irrepresentable conditions for lasso estimates when all groups correspond to singletons.

3.2. Group Restricted Eigenvalue Condition and Group Compatibility Condition. Restricted eigenvalue conditions [Bickel, Ritov and Tsybakov, 2009] ensure minimax optimal ℓ_2 estimation error in several penalized regression problems van de Geer and Bühlmann [2009], while the analogue for group lasso problems is introduced in Lounici et al. [2011]. In the regression framework of (A.1), RE(s, L) is satisfied, if there exists a positive number $\phi_{RE} = \phi_{RE}(s) > 0$ such that

$$(3.10) \quad \min_{\substack{J \subset \mathbb{N}_G, |J| \leq s \\ \Delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}}} \left\{ \frac{\|X \Delta\|}{\sqrt{n} \|\Delta_{[J]}\|} : \sum_{g \in J^c} \lambda_g \|\Delta_{[g]}\| \leq L \sum_{g \in J} \lambda_g \|\Delta_{[g]}\| \right\} \geq \phi_{RE}$$

Oracle inequalities for consistency of group lasso estimators in $\ell_{2,1}$ norms under a RE(s, 3) assumption and consistency in ℓ_2 norms under an RE(2s, 3) assumption are discussed in Lounici et al. [2011].

Following van de Geer and Bühlmann [2009], we introduce a slightly weaker notion called **Group Compatibility** (GC). For a constant $L > 0$ we say that GC(S, L) condition holds, if there exists a constant

$\phi_{compatible} = \phi_{compatible}(S, L) > 0$ such that
(3.11)

$$\min_{\Delta \in \mathbb{R}^p \setminus \{0\}} \left\{ \frac{\left(\sum_{g \in S} \lambda_g^2 \right)^{1/2} \|X\Delta\|}{\sqrt{n} \sum_{g \in S} \lambda_g \|\Delta_{[g]}\|} : \sum_{g \notin S} \lambda_g \|\Delta_{[g]}\| \leq L \sum_{g \in S} \lambda_g \|\Delta_{[g]}\| \right\} \geq \phi_{compatible}$$

This notion is used to connect the irrerepresentable conditions to the consistency results of group lasso estimators in $\ell_{2,1}$ norms. The fact that GC(S, L) holds whenever RE(s, L) is satisfied follows directly from the Cauchy Schwarz inequality.

4. Main Results. As discussed earlier, a number of authors have investigated the norm consistency of generic group lasso estimates under different assumptions, and asymptotic regimes [Bach, 2008, Nardi and Rinaldo, 2008, Wei and Huang, 2010, Lounici et al., 2011]. In particular, Lounici et al. [2011] establish the norm consistency of group lasso estimates under restricted eigenvalue assumptions. Of main interest, is to derive conditions that establish the validity of these assumptions in the context of NGC models. This issue is addressed in Sections 4.1 and 4.2. Subsequently, employing the notion of direction consistency introduced in Section 3.1, we establish selection consistency of the generic group lasso estimate, and investigate both the group-level and within group consistency of thresholded group lasso estimates for NGC.

4.1. Norm consistency of generic group lasso estimates. We start by presenting for the NGC framework independent derivations of the results established in Lounici et al. [2011], under slightly different choices of tuning parameters and assumptions. Asymptotically both estimates share the same convergence rate. However, we use a compatibility condition analogous to the one in van de Geer and Bühlmann [2009], instead of $RE(s, 3)$ assumption of Lounici et al. [2011], to derive finite sample estimation error bounds in the $\ell_{2,1}$ norm.

PROPOSITION 4.1. *Suppose the GC condition (3.11) holds with $L = 3$. Choose $\alpha > 0$ and denote $\lambda_{min} = \min_{1 \leq g \leq G} \lambda_g$. If*

$$\lambda_g \geq \frac{2\sigma}{\sqrt{n}} \sqrt{\|C_{[g][g]}\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right)$$

for every $g \in \mathbb{N}_G$, then, the following statements hold with probability at

least $1 - 2G^{1-\alpha}$,

$$(4.1) \quad \frac{1}{n} \left\| X \left(\hat{\beta} - \beta^0 \right) \right\|^2 \leq \frac{16}{\phi_{\text{compatible}}^2} \sum_{g=1}^s \lambda_g^2$$

$$(4.2) \quad \|\hat{\beta} - \beta^0\|_{2,1} \leq \frac{16}{\phi_{\text{compatible}}^2} \frac{\sum_{g=1}^s \lambda_g^2}{\lambda_{\min}}.$$

If, in addition, $RE(2s, 3)$ holds, then, with the same probability we get

$$(4.3) \quad \|\hat{\beta} - \beta^0\| \leq \frac{4\sqrt{10}}{\phi_{RE}^2(2s)} \frac{\sum_{g=1}^s \lambda_g^2}{\lambda_{\min} \sqrt{s}}.$$

The result shows that group lasso achieves faster convergence rate than lasso, if the groups are appropriately specified. Note that if all groups are of equal size k and $\lambda_g = \lambda$ for all g , then group lasso has an ℓ_2 estimation error of order $O\left(\sqrt{s}(\sqrt{k} + \sqrt{\log G})/\sqrt{n}\right)$. In contrast, lasso's error is $\sqrt{\|\beta^0\|_0 \log p/n}$, which establishes that group lasso has a lower error bound if $s \ll \|\beta^0\|_0$. On the other hand, lasso will have a lower error bound if $s \asymp \|\beta^0\|_0$, i.e., if the groups are highly misspecified.

Next, we investigate when the restricted eigenvalue and compatibility conditions hold. [Raskutti, Wainwright and Yu \[2010\]](#), [Rudelson and Zhou \[2011\]](#) discuss the RE assumption for lasso for different families of random design matrices and error distributions. In particular, [Raskutti, Wainwright and Yu \[2010\]](#) show that the restricted eigenvalue condition for lasso holds with high probability if the sample size is large enough ($n \asymp q \log p$) and the minimum eigenvalue of the covariance matrix of each row of the design matrix (i.e. $\Lambda_{\min}(\Sigma)$) is bounded away from 0. The following is an adaptation of that result, tailored to group lasso regression.

PROPOSITION 4.2. *Consider a generic group lasso regression (3.1) with a Gaussian random design matrix $X \in \mathbb{R}^{n \times p}$ whose rows are i.i.d. $N(\mathbf{0}, \Sigma)$. If $\Sigma^{1/2}$ satisfies $RE(s, 3)$ with a constant ϕ_{RE} (which holds trivially if $\Lambda_{\min}(\Sigma) > 0$), then there exist universal positive constants c, c', c'' , such that if the sample size n satisfies*

$$n > c'' \frac{16\rho^2(\Sigma)}{\phi_{RE}^2} \left(\frac{s(\sqrt{\log G} + \sqrt{k_g})^2}{\lambda_{\min}/\lambda_{\max}} \right), \quad \text{where } \rho^2(\Sigma) = \max_{1 \leq g \leq G} \|\Sigma_{[g][g]}\|$$

then X also satisfies $RE(s, 3)$ with $\phi_{RE}/8$ with probability at least $1 - c' \exp(-cn)$.

4.2. *Norm Consistency of Group NGC estimates.* In view of the above results, norm consistency of the regular group NGC estimates holds under an appropriate asymptotic regime, if both the restricted eigenvalue and group compatibility conditions are satisfied with high probability. The following result, together with Proposition 4.2, achieves this objective. Specifically, it shows that for a regular NGC estimation problem (2.3), $\Lambda_{\min}(\Sigma)$ is bounded away from 0, as long as the underlying VAR model is stable [cf. Lütkepohl, 2005], with its cross spectral density and the true adjacency matrices bounded above in spectral norm.

PROPOSITION 4.3. *Consider a stable, stationary VAR(d) model of the form (2.1). Let $\Sigma = \text{Var}(\mathbf{X}^{1:T})$ and $f(\theta)$, $\theta \in [-\pi, \pi]$ denote its cross spectral density. Suppose the spectral norm of the characteristic polynomial $A(z) = I - A^1 z - A^2 z^2 - \dots - A^d z^d$ evaluated on the circle $|z| = 1$ is bounded above, i.e., $\exists M > 0$ such that $\|A(e^{-i\theta})\| < M$, $\theta \in [-\pi, \pi]$. Then $\Lambda_{\min}(\Sigma) > \frac{1}{M}$. In particular this is satisfied when $m := \max_{1 \leq t \leq d} \|A^t\| < \infty$, for some $m > 0$.*

COROLLARY 4.4. *If the maximum incoming and outgoing effects at every node are bounded above, i.e., if*

$$(4.4) \quad \mathbf{v}_{in} = \max_{1 \leq i \leq p} \sum_{t=1}^d \sum_{j=1}^p |A_{ij}^t| < \infty, \quad \mathbf{v}_{out} = \max_{1 \leq j \leq p} \sum_{t=1}^d \sum_{i=1}^p |A_{ij}^t| < \infty$$

then $\Lambda_{\min}(\Sigma)$ is bounded away from 0.

PROOF. This corollary is a simple consequence of the above proposition together with the following result relating different norms for a matrix, [see e.g. Golub and Van Loan, 1996, Cor 2.3.2],

$$\|A^t\|_2 \leq \sqrt{\|A^t\|_1 \|A^t\|_\infty} \leq \frac{\|A^t\|_1 + \|A^t\|_\infty}{2} \quad t = 1, \dots, d$$

and the definitions

$$\|A^t\|_1 = \max_{1 \leq i \leq p} \sum_{j=1}^p |A_{ij}^t|, \quad \|A^t\|_\infty = \max_{1 \leq j \leq p} \sum_{i=1}^p |A_{ij}^t|$$

□

The following theorem is an immediate corollary of the above results.

THEOREM 4.5. *Consider a NGC estimation problem (2.3). Suppose the common design matrix $\mathbf{X}^n = [\mathcal{X}^1 : \dots : \mathcal{X}^{T-1}]$ in the p regression problems satisfy $RE(2s, 3)$ with $s = \max_i |pa_i|$, where pa_i denotes the set of parent nodes of X_i^T in the network. Consider the asymptotic regimes $G \asymp n^a$, $a > 0$ and $s = O(n^{c_1})$, $k_{\max} = O(n^{c_2})$, $0 < c_1, c_2 < 1$ such that $\sqrt{s}(\sqrt{k_{\max}} + \sqrt{\log G})/\sqrt{n} = o(1)$. Then for a suitably chosen sequence of λ_n we have $\|\hat{A}^{1:\hat{d}} - A^{1:d}\|_F \rightarrow 0$ in probability, as $n, p \rightarrow \infty$.*

4.3. Selection consistency for generic group lasso estimates. Next, we discuss the selection consistency properties of a generic group lasso regression problem with a common tuning parameter across groups, i.e., $\lambda_g = \lambda$ for every $g \in \mathbb{N}_G$. Similar results can be obtained for more general choices of the tuning parameters.

THEOREM 4.6. *Assume that the group uniform irrepresentable condition holds with $1 - \eta$ for some $\eta > 0$. Then, for any choice of*

$$\begin{aligned} \lambda &\geq \max_{g \notin S} \frac{1}{\eta} \frac{\sigma}{\sqrt{n}} \sqrt{\|(C_{22})_{[g][g]}\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right) \\ \delta_n &\geq \max_{g \in S} \frac{1}{\|\beta_{[g]}^0\|} \left(\lambda \sqrt{s} \|(C_{11})^{-1}\| + \sigma \sqrt{\|(C_{11})_{[g][g]}^{-1}\|} \frac{(\sqrt{k_g} + \sqrt{\alpha \log G})}{\sqrt{n}} \right), \end{aligned}$$

with probability greater than $1 - 4G^{1-\alpha}$, there exists a solution $\hat{\beta}$ satisfying

1. $\hat{\beta}_{[g]} = 0$ for all $g \notin S$,
2. $\|\hat{\beta}_{[g]} - \beta_{[g]}^0\| < \delta_n \|\beta_{[g]}^0\|$, and hence $\|D(\hat{\beta}_{[g]}) - D(\beta_{[g]}^0)\| < 2\delta_n$, for all $g \in S$. If $\delta_n < 1$, then $\hat{\beta}_{[g]} \neq 0$ for all $g \in S$.

REMARK 4.7. *The tuning parameter λ can be chosen of the same order as required for ℓ_2 consistency to achieve selection consistency within groups in the sense of (3.3). Further, with the above choice of λ , δ_n can be chosen of the order of $O(\sqrt{s}(\sqrt{k_{\max}} + \sqrt{\log G})/\sqrt{n})$. Thus, group lasso correctly identifies the group sparsity pattern if $\sqrt{s}(\sqrt{k_{\max}} + \sqrt{\log G})/\sqrt{n} \rightarrow 0$, the same scaling required for ℓ_2 consistency.*

Note that, the second part of the Theorem 4.6 also shows that group lasso estimates are direction consistent under the same scaling and hence a thresholded version of the estimates selects all important variables with high probability, as discussed in section 4.4. It can be shown that the weak irrepresentable condition is necessary for direction consistency of the group

lasso estimates under mild regularity conditions on the design matrices. In addition, analogously to the result in [van de Geer and Bühlmann, 2009], it can be shown that a slightly stronger version of the uniform irrerepresentable condition implies group compatibility conditions for group lasso estimates. We refer to Appendix D for a detailed discussion of these connections.

4.4. *Thresholding in Group NGC estimators.* As described in Section 2.2, regular group NGC estimates can be thresholded both at the group and coordinate levels. The first level of thresholding is motivated by the fact that lasso can select too many false positives [cf. van de Geer, Bühlmann and Zhou [2011], Zhou [2010] and the references therein]. We propose a hard-thresholding of regular group NGC estimates using a threshold $\delta_{grp} = C\lambda$ for some suitably chosen constant C . The second level of thresholding employs the direction consistency of regular group NGC estimates to perform within group variable selection with high probability. At this level, we hard-threshold a coordinate $j \in \mathcal{G}_g$ to zero if the corresponding coordinate of $D(\hat{\beta}_{[g]})$ is lower than a threshold $\delta_n \in (0, 1)$ in absolute value. In view of Theorem 4.6, the within group thresholding selects the group members with strong enough signal relative to other members of that group. The following result demonstrates the benefit of these two types of thresholding. Note that the thresholding at group level relies only on a weak GC(S, 3) condition, while the within group thresholding requires a stronger irrerepresentable condition.

THEOREM 4.8. *Consider a generic group lasso regression problem (3.1) with common tuning parameter $\lambda_g = \lambda$.*

- i) *Assume the GC(S, 3) condition of (3.11) holds with a constant $\phi = \phi_{compatible}$ and define*

$$\hat{\beta}_{[g]}^{thgrp} = \hat{\beta}_{[g]} \mathbf{1}_{\|\hat{\beta}_{[g]}\| > 4\lambda}.$$

If $\hat{S} = \{g \in \mathbb{N}_G : \hat{\beta}_{[g]}^{thgrp} \neq \mathbf{0}\}$, then $|\hat{S} \setminus S| \leq \frac{s}{\phi^2/12}$, with probability at least $1 - 2G^{1-\alpha}$.

- ii) *Assume that uniform irrerepresentable condition holds with $1 - \eta$ for some $\eta > 0$. Choose λ and δ_n as in Theorem 4.6 and define*

$$\hat{\beta}_j^{thgrp} = \hat{\beta}_j \mathbf{1}_{\{|\hat{\beta}_j| / \|\hat{\beta}_{[g]}\| > 2\delta_n\}} \text{ for all } j \in \mathcal{G}_g$$

Then, we have $\text{supp}(\beta^0) = \text{supp}(\hat{\beta}^{thgrp})$ with probability at least $1 - 4G^{1-\alpha}$ if $\min_{j \in \text{supp}(\beta^0)} |\beta_j^0| > 2\delta_n \|\beta_{[g]}^0\|$ for all $j \in \mathcal{G}_g$, i.e., the effect of every non-zero member in a group is “visible” relative to the total effect from the group.

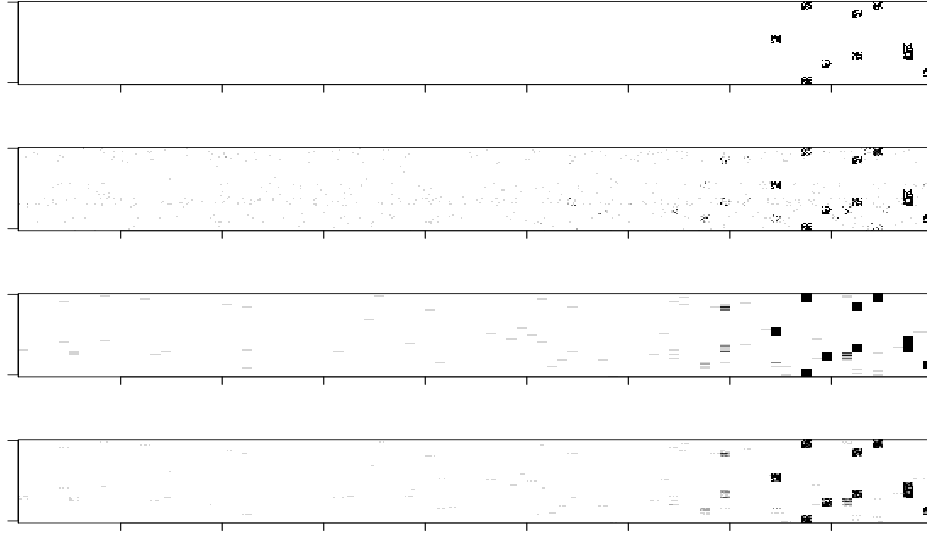


Fig 2: Estimated adjacency matrices of a misspecified NGC model: (a) True, (b) Lasso, (c) Group Lasso, (d) Thresholded Group Lasso

In NGC settings, where information about the temporal decay of the edge density of the network is available, a third level of thresholding is useful. Specifically, one can shrink to zero all the coefficients for each time lag where the total number of edges does not exceed a prespecified threshold that takes into account a predefined probability for false negatives. In this work, we do not further pursue such estimators.

5. Performance Evaluation. We evaluate the performances of regular, adaptive and thresholded variants of the group NGC estimators through an extensive simulation study, and compare the results to those obtained from lasso estimates. A standard R package (`grpreg` [Breheny and Huang, 2009]) was used to obtain the estimates.

The settings considered are:

1. *Balanced groups of equal size*: The parameters are as follows: i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 5$ time points, comprising of $p = 60, 120, 200$ nodes partitioned into groups of equal size in the range 3-5.
2. *Unbalanced groups*: In this case, the corresponding node set is partitioned into one larger group of size 10 and many groups of size 5.

3. *Misspecified balanced groups*: The parameters are as follows: i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 10$ time points, comprising of $p = 60, 120$ nodes partitioned into groups of equal size 6. Further, for each group there is a 30% misspecification rate, namely that for every parent group of a downstream node, 30% of the group members do not exert any effect on it.

The choice of the best tuning parameter λ is based on a grid search in the interval $[C_1\lambda_e, C_2\lambda_e]$ where $\lambda_e = \sqrt{2 \log p/n}$ for lasso and $\sqrt{2 \log G/n}$ for group lasso, using a 19 : 1 sample-splitting. The thresholding parameters are selected as $\delta_{grp} = 0.7\lambda\sigma$ at the group level and $\delta_{misspec} = n^{-0.2}$ within groups. Finally, within group thresholding is applied only when the group structure is misspecified.

The following performance metrics were used for comparison purposes: *Precision* = $TP/(TP + FP)$, (ii) *Recall* = $TP/(TP + FN)$ and (iii) Matthew's Correlation coefficient (MCC) defined as

$$\frac{(TP \times TN) - (FP \times FN)}{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{1/2}}$$

where TP , TN , FP and FN correspond to true positives, true negatives, false positives and false negatives in the estimated network, respectively.

The results for the balanced settings are given in Table 1. The average and standard deviations (in parentheses) of the performance metrics are presented for each setup. The Recall for $p = 60$ shows that even for a network with $60 \times (5 - 1) = 240$ nodes and $|E| = 351$ true edges, the group NGC estimators recover about 71% of the true edges with a sample size as low as $n = 60$, while lasso based NGC estimates recover only 31% of the true edges. The three group NGC estimates have comparable performances in all the cases. However thresholded lasso shows slightly higher precision than the other group NGC variants for smaller sample sizes (e.g., $n = 60, p = 200$). The results for $p = 60, n = 110$ also display that lower precision of lasso is caused partially by its inability to estimate the order of the VAR model correctly, as measured by ERR LAG=Number of falsely connected edges from lags beyond the true order of the VAR model divided by the number of edges in the network ($|E|$). This finding is nicely illustrated in Figure 2 and Table 1. The group penalty encourages edges from the nodes of the same group to be picked up together. Since the nodes of the same group are also from the same time lag, the group variants have substantially lower ERR LAG. For example, average ERR LAG of lasso for $p = 200, n = 160$

is 19.79% while the average ERR LAGs for the group lasso variants are in the range 3.06% – 4.21%.

The results for the unbalanced networks are given in Table 2. As in the balanced group setup, in almost all the simulation settings the group NGC variants outperform the lasso estimates with respect to all three performance metrics. However the performances of the different variants of group NGC are comparable and tend to have higher standard deviations than the lasso estimates. Also the average ERR LAGs for the group NGC variants are substantially lower than the average ERR LAG for lasso demonstrating the advantage of group penalty. Although the conclusions regarding the comparisons of lasso and group NGC estimates remain unchanged it is evident that the performances of all the estimators are affected by the presence of one large group, skewing the uniform nature of the network. For example the MCC measures of group NGC estimates in a balanced network with $p = 60$ and $|E| = 351$ vary around 97 – 98% which lowers to 89% – 90% when the groups are unbalanced.

The results for misspecified groups are given in Table 3. Note that for higher sample size n the MCC of lasso and regular group lasso are comparable. However, the thresholded version of group lasso ($\delta_{misspec} = n^{-0.2}$ used for within group selection) achieves significantly higher MCC than the rest. This demonstrates the advantage of using the directional consistency of group lasso estimators to perform within group variable selection. We would like to mention here that a careful choice of the thresholding parameters δ_{grp} and $\delta_{misspec}$ via cross-validation or other model selection criteria indicate improvement in the performance of thresholded group lasso; however, we do not pursue these methods here as they require grid search over many tuning parameters or an efficient estimator of the degree of freedom of group lasso.

In summary, the results clearly show that all variants of group lasso NGC outperform the lasso-based ones, whenever the grouping structure of the variables is known and correctly specified. Further, their performance depends on the composition of group sizes. On the other hand, if the a priori known group structure is moderately misspecified lasso estimates produce comparable results to regular and adaptive group NGC ones, while thresholded group estimates outperform all other methods, as expected.

6. Application.

6.1. *Example: Banking balance sheets application.* In this application, we examine the structure of the balance sheets in terms of assets and liabilities of the $n = 50$ largest (in terms of total balance sheet size) US banking

corporations. The data cover 9 quarters (September 2009-September 2011) and were directly obtained from the Federal Deposit Insurance Corporation (FDIC) database (available at www.fdic.gov). The $p = 21$ variables correspond to different assets (US and foreign government debt securities, equities, loans (commercial, mortgages), leases, etc.) and liabilities (domestic and foreign deposits from households and businesses, deposits from the Federal Reserve Board, deposits of other financial institutions, non-interest bearing liabilities, etc.) We have organized them into four categories: two for the assets (loans and securities) and two for the liabilities (Balances Due and Deposits, based on a \$250K reporting FDIC threshold). Amongst the 50 banks examined, one discerns large integrated ones with significant retail, commercial and investment activities (e.g. Citibank, JP Morgan, Bank of America, Wells Fargo), banks primarily focused on investment business (e.g. Goldman Sachs, Morgan Stanley, American Express, E-Trade, Charles Schwab), regional banks (e.g. Banco Popular de Puerto Rico, Comerica Bank, Bank of the West).

The raw data are reported in thousands of dollars. The few missing values were imputed using a nearest neighbor imputation method with $k = 5$, by clustering them according to their total assets in the most recent quarter (September 2011) and subsequently every missing observation for a particular bank was imputed by the median observation on its five nearest neighbors. The data were log-transformed to reduce non-stationarity issues. The dataset was restructured as a panel with $p = 21$ variables and $n = 50$ replicates observed over $T = 9$ time points. Every column of replicates was scaled to have unit variance.

We applied the proposed variants of NGC estimates on the first $T = 6$ time points (Sep 2009 - Dec 2010) of the above panel dataset. The parameters λ and δ_{grp} were chosen using a 19 : 1 sample-splitting method and the misspecification threshold $\delta_{misspec}$ was set to zero as the grouping structure was reliable. We calculated the MSE of the fitted model in predicting the outcomes in the four quarters (December 2010 - September 2011). The Predicted MSE (MSE for Dec 2010) are listed in Table 4. The estimated network structures are shown in Figures 3 and 4.

It can be seen that the lasso estimates recover a very simple temporal structure amongst the variables; namely, that past values (in this case lag-1) influence present ones. Given the structure of the balance sheet of large banks, this is an anticipated result, since it can not be radically altered over a short time period due to business relationships and past commitments to customers of the bank. However, the (adaptive) group lasso estimates reveal a richer and more nuanced structure. Examining the fitted values

of the adjacency matrices A^t , we notice that the dominant effects remain those discovered by the lasso estimates. However, fairly strong effects are also estimated within each group, but also between the groups of the assets (loans and securities) on the balance sheet. This suggests rebalancing of the balance sheet for risk management purposes between relatively low risk securities and potentially more risky loans. Given the period covered by the data (post financial crisis starting in September 2009) when credit risk management became of paramount importance, the analysis picks up interesting patterns. On the other hand, significant fewer associations are discovered between the liabilities side of the balance sheet. Finally, there exist relationships between deposits and securities such as US Treasuries and other domestic ones (primarily municipal bonds); the latter indicates that an effort on behalf of the banks to manage the credit risk of their balance sheets, namely allocating to low risk assets as opposed to more risky loans.

It is also worth noting that the group lasso model exhibits superior predictive performance over the lasso estimates, even 4 quarters into the future. Finally, in this case the thresholded estimates did not provide any additional benefits over the regular and adaptive variants, given that the specification of the groups was based on accounting principles and hence correctly structured.

6.2. Example: T-cell activation. Estimation of gene regulatory networks from expression data is a fundamental problem in functional genomics [Friedman, 2004]. Time course data coupled with NGC models are informationally rich enough for the task at hand. The data for this application come from Rangel et al. [2004], where expression patterns of genes involved in T-cell activation were studied with the goal of discovering regulatory mechanisms that govern them in response to external stimuli. Activated T-cells are involved in regulation of effector cells (e.g. B-cells) and play a central role in mediating immune response. The available data comprising of $n = 44$ samples of $p = 58$ genes, measure the cells response at 10 time points, $t = 0, 2, 4, 6, 8, 18, 24, 32, 48, 72$ hours after their stimulation with a T-cell receptor independent activation mechanism. We concentrate on data from the first 5 time points, that correspond to early response mechanisms in the cells.

Genes are often grouped based on their function and activity patterns into biological pathways. Thus, the knowledge of gene functions and their membership in biological pathways can be used as inherent grouping structures in the proposed group lasso estimates of NGC. Towards this, we used available biological knowledge to define groups of genes based on their bi-

ological function. Reliable information for biological functions were found from the literature for 38 genes, which were retained for further analysis. These 38 genes were grouped into 13 groups with the number of genes in different groups ranging from 1 to 5.

In [Shojaie, Basu and Michailidis \[2012\]](#), we analyzed this data and showed that the decay condition for the truncating lasso penalty seems to be violated in this case, and considered instead estimation of regulatory effect using an adaptive thresholding penalty. Hence, we consider here only application of the adaptive and thresholding variants of the proposed group lasso estimator for NGC. Figure 5 shows the estimated networks based on lasso and thresholded group lasso estimates, where for ease of representation the nodes of the network represent groups of genes.

In this case, estimates from variants of group NGC estimator were all similar, and included a number of known regulatory mechanisms in T-cell activation, not present in the regular lasso estimate. For instance, [Waterman et al. \[1990\]](#) suggest that TCF plays a significant role in activation of T-cells, which may describe the dominant role of this group of genes in the activation mechanism. On the other hand, [Kim et al. \[2005\]](#) suggest that activated T-cells exhibit high levels of osteoclast-associated receptor activity which may attribute the large number of associations between member of osteoclast differentiation and other groups. Finally, the estimated networks based on variants of group lasso estimator also offer improved estimation accuracy in terms of mean squared error (MSE) despite having having comparable complexities to their regular lasso counterpart (Table 5), which further confirms the findings of other numerical studies in the paper.

7. Discussion. In this paper, the problem of estimating Network Granger Causal (NGC) models with inherent grouping structure is studied when replicates are available. Norm, and both group level and within group variable selection consistency are established under fairly mild assumptions on the structure of the underlying time series. To achieve the second objective the novel concept of direction consistency is introduced.

The type of NGC models discussed in this study have wide applicability in different areas, including genomics and economics. However, in many contexts the availability of replicates at each time point is not feasible (e.g. in rate of returns for stocks or other macroeconomic variables), while grouping structure is still present (e.g. grouping of stocks according to industry sector). Hence, it is of interest to study the behavior of group lasso estimates in such a setting and address the technical challenges emanating from such a pure time series (dependent) data structure.

APPENDIX A: AUXILIARY LEMMAS

LEMMA A.1 (Characterization of the Group lasso estimate). *A vector $\hat{\beta} \in \mathbb{R}^p$ is a solution to the convex optimization problem*

$$(A.1) \quad \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\beta\|^2 + \sum_{g=1}^G \lambda_g \|\beta_{[g]}\|$$

if and only if $\hat{\beta}$ satisfies for some $\tau \in \mathbb{R}^p$ with $\max_g \tau_{[g]} \leq 1$,

$$\frac{1}{n} \left[X'(Y - X\hat{\beta}) \right]_{[g]} = \lambda_g \tau_{[g]},$$

and $\tau_{[g]} = D(\hat{\beta}_{[g]})$ whenever $\hat{\beta}_{[g]} \neq \mathbf{0}$.

PROOF. The result follows directly from the KKT conditions for the optimization problem (A.1). \square

LEMMA A.2. *Let $Z \sim N(0, \Sigma)$ be a k -dimensional centered Gaussian random variable. Then, for any $t > 0$, the following concentration inequality holds:*

$$\mathbb{P}[|\|Z\| - \mathbb{E}\|Z\|| > t] \leq 2 \exp\left(-\frac{2t^2}{\pi^2 \|\Sigma\|}\right)$$

Further, $\mathbb{E}\|Z\| \leq \sqrt{k} \sqrt{\|\Sigma\|}$.

PROOF. The first inequality can be found in [Ledoux and Talagrand \[1991\]](#) (equation (3.2)). To establish the second inequality note that,

$$\mathbb{E}\|Z\| \leq \sqrt{\mathbb{E}\|Z\|^2} = \sqrt{\mathbb{E}[\operatorname{tr}(ZZ')] } = \sqrt{\operatorname{tr}(\Sigma)} \leq \sqrt{k} \sqrt{\|\Sigma\|}$$

\square

LEMMA A.3. *Let $\beta, \hat{\beta} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$. Denote by $\hat{u} = \hat{\beta} - \beta$ and $r = D(\hat{\beta}) - D(\beta)$. Then, if $\|\hat{u}\| < \delta \|\beta\|$ we obtain $\|r\| < 2\delta$.*

PROOF. It follows from $\|\hat{u}\| < \delta \|\beta\|$, that

$$(1 - \delta)\|\beta\| < \|\beta\| - \|\hat{u}\| \leq \|\hat{\beta}\| \leq \|\hat{u}\| + \|\beta\| < (1 + \delta)\|\beta\|,$$

which implies that $\left| \|\beta\| - \|\hat{\beta}\| \right| < \delta \|\beta\|$. Now,

$$\begin{aligned} \|r\| &= \frac{1}{\|\hat{\beta}\| \|\beta\|} \left\| \hat{\beta} \|\beta\| + (\hat{u} - \hat{\beta}) \|\hat{\beta}\| \right\| \\ &\leq \frac{1}{\|\hat{\beta}\| \|\beta\|} \left\| \hat{\beta} (\|\beta\| - \|\hat{\beta}\|) + \|\hat{\beta}\| \hat{u} \right\| \\ &< \frac{1}{\|\hat{\beta}\| \|\beta\|} \left(\|\hat{\beta}\| \delta \|\beta\| + \|\hat{\beta}\| \|\hat{u}\| \right) < \delta + \delta = 2\delta \end{aligned}$$

□

APPENDIX B: RESULTS FOR NORM CONSISTENCY

PROOF OF PROPOSITION (4.1). Since $\hat{\beta}$ is a solution of the optimization problem (3.1), for all $\beta \in \mathbb{R}^p$, we have

$$\frac{1}{n} \|Y - X\hat{\beta}\|^2 + 2 \sum_{g=1}^G \lambda_g \|\hat{\beta}_{[g]}\| \leq \frac{1}{n} \|Y - X\beta\|^2 + 2 \sum_{g=1}^G \lambda_g \|\beta_{[g]}\|.$$

Plugging in $Y = X\beta^0 + \epsilon$, and simplifying the resulting equation, we get

$$\begin{aligned} \frac{1}{n} \|X(\hat{\beta} - \beta^0)\|^2 &\leq \frac{1}{n} \|X(\beta - \beta^0)\|^2 + \frac{2}{n} \sum_{g=1}^G \|(X'\epsilon)_{[g]}\| \|(\hat{\beta} - \beta)_{[g]}\| \\ &\quad + 2 \sum_{g=1}^G \lambda_g \left(\|\beta_{[g]}\| - \|\hat{\beta}_{[g]}\| \right). \end{aligned}$$

Fix $g \in \mathbb{N}_G$ and consider the event $\mathcal{A}_g = \left\{ \epsilon \in \mathbb{R}^n : \frac{2}{n} \|(X'\epsilon)_{[g]}\| \leq \lambda_g \right\}$. Note that $Z = \frac{1}{\sqrt{n}} X'\epsilon \sim N(\mathbf{0}, \sigma^2 C)$. So $Z_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{[g][g]})$. Then,

$$\begin{aligned} \mathbb{P}(\mathcal{A}_g^c) &= \mathbb{P}\left(\|Z_{[g]}\| > \frac{1}{2} \lambda_g \sqrt{n}\right) \\ &\leq \mathbb{P}\left(\|Z_{[g]} - \mathbb{E} \|Z_{[g]}\|\| > \frac{\lambda_g \sqrt{n}}{2} - \sigma \sqrt{k_g} \sqrt{\|C_{[g][g]}\|}\right), \end{aligned}$$

where the last inequality follows from the second statement of Lemma A.2.

Now, let $x_g = \frac{\lambda_g \sqrt{n}}{2} - \sigma \sqrt{k_g} \sqrt{\|C_{[g][g]}\|}$. Then, for $x_g > 0$, if

$$2 \exp\left(-\frac{2x_g^2}{\pi^2 \sigma^2 \|C_{[g][g]}\|}\right) \leq 2G^{-\alpha},$$

we get

$$\mathbb{P}(\mathcal{A}_g^c) \leq 2G^{-\alpha}.$$

But this happens if,

$$\sqrt{2}x_g \geq \sqrt{\alpha \log G} \pi \sigma \sqrt{\|C_{[g][g]}\|},$$

which is ensured by the proposed choice of λ_g .

Next, define $\mathcal{A} := \cap_{g=1}^G \mathcal{A}_g$. Then, $\mathbb{P}(\mathcal{A}) \geq 1 - 2G^{1-\alpha}$, and on the event \mathcal{A} , we have, for all $\beta \in \mathbb{R}^p$,

$$\begin{aligned} \frac{1}{n} \|X(\hat{\beta} - \beta^0)\|^2 + \sum_{g=1}^G \lambda_g \left\| \hat{\beta}_{[g]} - \beta_{[g]} \right\| &\leq \frac{1}{n} \|X(\beta - \beta^0)\|^2 \\ &+ 2 \sum_{g=1}^G \lambda_g \left(\left\| \hat{\beta}_{[g]} - \beta_{[g]} \right\| + \left\| \beta_{[g]} \right\| - \left\| \hat{\beta}_{[g]} \right\| \right). \end{aligned}$$

Note that $\left(\left\| \hat{\beta}_{[g]} - \beta_{[g]} \right\| + \left\| \beta_{[g]} \right\| - \left\| \hat{\beta}_{[g]} \right\| \right)$ vanishes if $g \notin S$ and is bounded above by $\min\{2 \left\| \beta_{[g]} \right\|, 2 \left(\left\| \beta_{[g]} \right\| - \left\| \hat{\beta}_{[g]} \right\| \right)\}$ if $g \in S$.

This leads to the following sparsity oracle inequality, for all $\beta \in \mathbb{R}^p$,

$$\begin{aligned} \frac{1}{n} \|X(\hat{\beta} - \beta^0)\|^2 + \sum_{g=1}^G \lambda_g \left\| \hat{\beta}_{[g]} - \beta_{[g]} \right\| &\leq \frac{1}{n} \|X(\beta - \beta^0)\|^2 \\ \text{(B.1)} \quad &+ 4 \sum_{g \in S} \lambda_g \min \left\{ \left\| \beta_{[g]} \right\|, \left\| \beta_{[g]} \right\| - \left\| \hat{\beta}_{[g]} \right\| \right\}. \end{aligned}$$

The sparsity oracle inequality (B.1) with $\beta = \beta^0$, and $\Delta := \hat{\beta} - \beta^0$ leads to the following two useful bounds on the prediction and $\ell_{2,1}$ -estimation errors:

$$\text{(B.2)} \quad \frac{1}{n} \|X\Delta\|^2 \leq 4 \sum_{g \in S} \lambda_g \left\| \Delta_{[g]} \right\|$$

$$\text{(B.3)} \quad \sum_{g \notin S} \lambda_g \left\| \Delta_{[g]} \right\| \leq 3 \sum_{g \in S} \lambda_g \left\| \Delta_{[g]} \right\|.$$

Now, assume the group compatibility condition 3.11 holds. Then,

$$\text{(B.4)} \quad \frac{1}{n} \|X\Delta\|^2 \leq 4 \sum_{g \in S} \lambda_g \left\| \Delta_{[g]} \right\| \leq \sqrt{\sum_{g \in S} \lambda_g^2} \frac{X\Delta}{\sqrt{n}} \frac{1}{\phi_{compatible}},$$

which implies the first inequality of proposition 4.1. The second inequality follows from

$$\begin{aligned} \lambda_{\min} \left\| \hat{\beta} - \beta \right\|_{2,1} &\leq \sum_{g=1}^G \lambda_g \left\| \Delta_{[g]} \right\| \leq 4 \sum_{g \in S} \lambda_g \left\| \Delta_{[g]} \right\| \\ &\leq 4 \sqrt{\sum_{g \in S} \lambda_g^2} \frac{\|X\Delta\|}{\sqrt{n}} \frac{1}{\phi_{\text{compatible}}} \leq \frac{16}{\phi_{\text{compatible}}^2} \sum_{g \in S} \lambda_g^2, \end{aligned}$$

where the last step uses (B.4).

The proof of the last inequality of proposition 4.1, i.e., the upper bound on ℓ_2 estimation error under $RE(2s)$, is the same as in Theorem 3.1 in Lounici et al. [2011] and is omitted. \square

PROOF OF PROPOSITION 4.3. We note that Σ is a $pT \times pT$ block Toeplitz matrix with $(i, j)^{th}$ block $(\Sigma_{ij})_{1 \leq i, j \leq T} := \Gamma(i - j)$, where $\Gamma(\ell)_{p \times p}$ is the autocovariance function of lag ℓ for the zero-mean VAR(d) process (2.1), defined as

$$(B.5) \quad \Gamma(\ell) = \mathbb{E}[\mathbf{X}^t (\mathbf{X}^{t-\ell})']$$

We consider the cross spectral density of the VAR(d) process (2.1)

$$(B.6) \quad f(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\theta}, \quad \theta \in [-\pi, \pi]$$

From standard results of spectral theory we know that $\Gamma(\ell) = \int_{-\pi}^{\pi} e^{i\ell\theta} f(\theta) d\theta$, for every ℓ .

We want to find a lower bound on the minimum eigenvalue of Σ , i.e., $\inf_{\|x\|=1} x' \Sigma x$. Consider an arbitrary pT -variate unit norm vector x formed by stacking the p -tuples x^1, \dots, x^T .

For every $\theta \in [-\pi, \pi]$ define $G(\theta) = \sum_{t=1}^T x^t e^{-it\theta}$ and note that

$$\begin{aligned} \int_{-\pi}^{\pi} G^*(\theta) G(\theta) d\theta &= \sum_{t=1}^T \sum_{\tau=1}^T (x^t)' (x^\tau) \int_{-\pi}^{\pi} e^{i(t-\tau)\theta} d\theta \\ &= \sum_{t=1}^T \sum_{\tau=1}^T (x^t)' (x^\tau) (2\pi \mathbf{1}_{\{t=\tau\}}) \\ &= 2\pi \sum_{t=1}^T (x^t)' (x^t) = 2\pi \|x\|^2 = 2\pi \end{aligned}$$

Also let $\mu(\theta)$ be the minimum eigenvalue of the Hermitian matrix $f(\theta)$. Following [Parter \[1961\]](#) we have the result

$$\begin{aligned}
x' \Sigma x &= \sum_{t=1}^T \sum_{\tau=1}^T (x^t)' \Gamma(t - \tau) x^\tau \\
&= \sum_{t=1}^T \sum_{\tau=1}^T (x^t)' \left(\int_{-\pi}^{\pi} e^{i(t-\tau)\theta} f(\theta) d\theta \right) x^\tau \\
&= \int_{-\pi}^{\pi} \left(\sum_{t=1}^T (x^t)' e^{it\theta} \right) f(\theta) \left(\sum_{\tau=1}^T x^\tau e^{-i\tau\theta} \right) d\theta \\
&= \int_{-\pi}^{\pi} G^*(\theta) f(\theta) G(\theta) d\theta \\
&\geq \int_{-\pi}^{\pi} \mu(\theta) (G^*(\theta) G(\theta)) d\theta \\
&\geq \left(\min_{\theta \in (-\pi, \pi)} \mu(\theta) \right) \int_{-\pi}^{\pi} G^*(\theta) G(\theta) d\theta = 2\pi \min_{\theta \in (-\pi, \pi)} \mu(\theta)
\end{aligned}$$

So $\Lambda_{\min}(\Sigma) \geq 2\pi \min_{\theta \in (-\pi, \pi)} \mu(\theta)$.

If $A(z) = I - A^1 z - A^2 z^2 - \dots - A^d z^d$ is the (matrix-valued) characteristic polynomial of the VAR(d) model [\(2.1\)](#), then we have the following representation (see eqn (9.4.23), [Priestley \[1981\]](#)):

$$f(\theta) = \frac{1}{2\pi} \sigma^2 (A(e^{-i\theta}))^{-1} (A(e^{-i\theta}))^{-T}$$

Thus, $2\pi\mu(\theta) = 2\pi\Lambda_{\min}(f(\theta)) = 2\pi/\Lambda_{\max}(f(\theta)^{-1}) \geq 1/\|A(e^{-i\theta})\|$. But $\|A(e^{-i\theta})\| \leq 1 + \sum_{t=1}^d \|A^t\|$ for every $\theta \in [-\pi, \pi]$. So the minimum eigenvalue of Σ is bounded away from zero as long as the spectral norms of the adjacency matrices are bounded above. □

APPENDIX C: RESULTS FOR SELECTION CONSISTENCY

PROOF OF THEOREM [4.6](#). Consider any solution $\hat{\beta}_R \in \mathbb{R}^q$ of the restricted group lasso problem

$$(C.1) \quad \operatorname{argmin}_{\beta \in \mathbb{R}^q} \frac{1}{2n} \|\mathbf{Y} - X_{(1)}\beta\|_2^2 + \lambda \sum_{g=1}^s \|\beta_{[g]}\|_2$$

and set $\hat{\beta} = [\hat{\beta}_R' : \mathbf{0}_{1 \times (p-q)}]'$. We show that such an augmented vector $\hat{\beta}$ satisfies the statements of Theorem [4.6](#) with high probability.

Let $\hat{u} = \hat{\beta}_{(1)} - \beta_{(1)}^0 = \hat{\beta}_R - \beta_{(1)}^0$. In view of lemmas A.1 and A.3, it suffices to show that the following events happen with probability at least $1 - 4G^{1-\alpha}$:

$$(C.2) \quad \|\hat{u}_{[g]}\| < \delta_n \|\beta_{[g]}^0\|, \text{ for all } g \in S$$

$$(C.3) \quad \frac{1}{n} \left\| [X'(\epsilon - X_{(1)}\hat{u})]_{[g]} \right\| \leq \lambda, \text{ for all } g \notin S$$

Note that, in view of Lemma A.1, $\hat{u} = (C_{11})^{-1} \left(\frac{1}{\sqrt{n}} Z_{(1)} - \lambda \tau \right)$ for some $\tau \in \mathbb{R}^q$ with $\|\tau_{[g]}\| \leq 1$ for all $g \in S$, and $Z = \frac{1}{\sqrt{n}} X' \epsilon = [Z'_{(1)} : Z'_{(2)}]'$. Thus, for any $g \in S$,

$$\begin{aligned} \mathbb{P} \left(\|\hat{u}_{[g]}\| > \delta_n \|\beta_{[g]}^0\| \right) &\leq \mathbb{P} \left(\left\| \left[(C_{11})^{-1} \left(\frac{1}{\sqrt{n}} Z_{(1)} - \lambda \tau \right) \right]_{[g]} \right\| > \delta_n \|\beta_{[g]}^0\| \right) \\ &\leq \mathbb{P} \left(\left\| \left[(C_{11})^{-1} Z_{(1)} \right]_{[g]} \right\| > \sqrt{n} \left[\delta_n \|\beta_{[g]}^0\| - \lambda \left\| \left[(C_{11})^{-1} \tau \right]_{[g]} \right\| \right] \right) \end{aligned}$$

Note that $V = (C_{11})^{-1} Z_{(1)} \sim N(\mathbf{0}, \sigma^2 (C_{11})^{-1})$. So $V_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{11}^{[g][g]})$, where $\Sigma^{[g][g]} := (\Sigma^{-1})_{[g][g]}$. Also, by the second statement of lemma A.2 we have $\mathbb{E} \|V_{[g]}\| \leq \sigma \sqrt{k_g} \sqrt{\|C_{11}^{[g][g]}\|}$.

Therefore,

$$\begin{aligned} &\mathbb{P} \left(\|\hat{u}_{[g]}\| > \delta_n \|\beta_{[g]}^0\| \right) \\ &\leq \mathbb{P} \left(\left| \|V_{[g]}\| - \mathbb{E} \|V_{[g]}\| \right| > \sqrt{n} \left[\delta_n \|\beta_{[g]}^0\| - \lambda \|(C_{11})^{-1}\| \sqrt{s} - \sigma \sqrt{k_g} \|C_{11}^{[g][g]}\| \right] \right) \\ &= 2 \exp \left[-\frac{2}{\pi^2 \|\sigma^2 C_{11}^{[g][g]}\|} \left(\sqrt{n} \delta_n \|\beta_{[g]}^0\| - \sqrt{n} \lambda \|(C_{11})^{-1}\| \sqrt{s} - \sigma \sqrt{k_g} \|C_{11}^{[g][g]}\| \right)^2 \right] \end{aligned}$$

For the proposed choice of δ_n , the above probability is bounded above by $2G^{-\alpha}$.

Next, for any $g \notin S$, we get

$$\begin{aligned} &\mathbb{P} \left(\frac{1}{n} \left\| [X'(\epsilon - X_{(1)}\hat{u})]_{[g]} \right\| > \lambda \right) \\ &\leq \mathbb{P} \left(\left\| [Z_{(2)} - C_{21} C_{11}^{-1} Z_{(1)}]_{[g]} \right\| > \sqrt{n} \lambda \left(1 - \left\| [C_{21} C_{11}^{-1} \tau]_{[g]} \right\| \right) \right) \end{aligned}$$

Defining $W = Z_{(2)} - C_{21} C_{11}^{-1} Z_{(1)} \sim N(\mathbf{0}, \sigma^2 (C_{22} - C_{21} C_{11}^{-1} C_{12}))$, the uniform irrerepresentable condition implies that the above probability is bounded above by $\mathbb{P}(\|W_{[g]}\| > \sqrt{n} \lambda \eta)$.

It can then be seen that $W_{[g]} \sim N(\mathbf{0}, \sigma^2 \bar{C}_{[g][g]})$, where $\bar{C} = C_{22} - C_{21}C_{11}^{-1}C_{12}$ denotes the Schur complement of C_{22} . As before, lemma A.2 establishes that

$$\begin{aligned} \mathbb{P}(\|W_{[g]}\| > \sqrt{n}\lambda\eta) &\leq \mathbb{P}\left(\left|\|W_{[g]}\| - \mathbb{E}\|W_{[g]}\|\right| > \sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right) \\ &\leq 2 \exp\left[-\frac{2}{\pi^2\|\sigma^2\bar{C}_{[g][g]}\|}\left(\sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right)^2\right], \end{aligned}$$

and the last probability is bounded above by $2G^{-\alpha}$ for the proposed choice of λ .

The results in the proposition follow by considering the union bound on the two sets of the probability statements made across all $g \in \mathbb{N}_G$. \square

PROOF OF THEOREM 4.8. We use the notations developed in the proof of Proposition 4.1. First note that, (ii) follows directly from Theorem 4.6. For (i), since the falsely selected groups are present after the initial thresholding, we get $\|\hat{\beta}_{[g]}\| > 4\lambda$ for every such group. Next, we obtain an upper bound for the number of such groups. Specifically, denoting $\Delta = \hat{\beta} - \beta^0$, we get

$$(C.4) \quad |\hat{S} \setminus S| \leq \frac{\|\hat{\beta}_{S^c}\|_{2,1}}{4\lambda} = \frac{\sum_{g \notin S} \|\Delta_{[g]}\|}{4\lambda}.$$

Next, note that from the sparsity oracle inequality (B.2), the following holds on the event \mathcal{A} ,

$$\sum_{g \notin S} \|\Delta_{[g]}\| \leq 3 \sum_{g \in S} \|\Delta_{[g]}\|$$

It readily follows that

$$4 \sum_{g \notin S} \|\Delta_{[g]}\| \leq 3 \|\Delta\|_{2,1} \leq \frac{48}{\phi^2} s\lambda$$

where the last inequality follows from the $\ell_{2,1}$ -error bound of (4.2). Using this inequality together with (C.4) gives the result. \square

APPENDIX D: SUPPLEMENTS

In this section, we discuss two results involving the compatibility and irrepresentable conditions for group lasso. The first result demonstrates a connection between irrepresentable conditions and compatibility conditions. The second result discusses the necessity of group irrepresentable conditions for direction consistency of the group lasso estimates. The proofs are given

under a special choice of tuning parameter $\lambda_g = \lambda\sqrt{k_g}$. Similar results can be derived for the general choice of λ_g , although their presentation is more involved.

The following result is a generalization of Theorem 9.1 in [van de Geer and Bühlmann \[2009\]](#).

PROPOSITION D.1. *Suppose uniform irrerepresentable condition (3.8) holds with $\eta \in [0, 1]$. Then group compatibility(S, L) (3.11) condition holds whenever $L < \frac{1}{1-\eta}$.*

PROOF. First note that with the above choice of λ_g the Group Compatibility (S, L) condition simplifies to
(D.1)

$$\phi_{compatible} := \min_{\Delta \in \mathbb{R}^p \setminus \{0\}} \left\{ \frac{\sqrt{q} \|X\Delta\|}{\sqrt{n} \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}\|} : \sum_{g \notin S} \sqrt{k_g} \|\Delta_{[g]}\| \leq L \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}\| \right\} > 0$$

Also, the uniform irrerepresentable condition guarantees that there exists $0 < \eta < 1$ such that $\forall \tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \leq g \leq s} \|\tau_{[g]}\|_2 \leq 1$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21} (C_{11})^{-1} K^0 \tau \right]_{[g]} \right\|_2 < 1 - \eta \quad \forall g \notin S$$

Here $K^0 = K/\lambda$ is a $q \times q$ block diagonal matrix with s diagonal blocks $\sqrt{k_1} \mathbf{I}_{k_1 \times k_1}, \dots, \sqrt{k_s} \mathbf{I}_{k_s \times k_s}$. Define
(D.2)

$$\Delta^0 := \operatorname{argmin}_{\Delta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{X}\Delta\|_2^2 : \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}\|_2 = 1, \sum_{g \notin S} \sqrt{k_g} \|\Delta_{[g]}\|_2 \leq L \right\}$$

Note that $\frac{1}{n} \|\mathbf{X}\Delta^0\|_2^2 = \phi_{compatible}^2/q$, and introduce two Lagrange multipliers λ and λ' corresponding to the equality and inequality constraints for solving the optimization problem in (D.2). Also, partition $\Delta^0 = [\Delta_{(1)}^0 : \Delta_{(2)}^0]$ and $\mathbf{X} = [\mathbf{X}_{(1)} : \mathbf{X}_{(2)}]$ into signal and nonsignal parts as in (3.6). The first q linear equations of the KKT conditions imply that there exists $\tau^0 \in \mathbb{R}^q$ such that

$$(D.3) \quad C_{11} \Delta_{(1)}^0 + C_{12} \Delta_{(2)}^0 = \lambda K^0 \tau^0$$

and, for every $g \in S$,

$$\begin{aligned}\tau_{[g]}^0 &= D(\Delta_{[g]}^0) \text{ if } \Delta_{[g]}^0 \neq \mathbf{0} \\ \|\tau_{[g]}^0\|_2 &\leq 1 \text{ if } \Delta_{[g]}^0 = \mathbf{0}\end{aligned}$$

It readily follows that $(\tau^0)^T K^0 \Delta_{(1)}^0 = \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}^0\|_2 = 1$.

Multiplying both sides of (D.3) by $(\Delta_{(1)}^0)^T$ we get

$$(D.4) \quad \left(\Delta_{(1)}^0\right)^T C_{11} \Delta_{(1)}^0 + \left(\Delta_{(1)}^0\right)^T C_{12} \Delta_{(2)}^0 = \lambda$$

Also, (D.3) implies

$$(D.5) \quad \Delta_{(1)}^0 + (C_{11})^{-1} C_{12} \Delta_{(2)}^0 = \lambda (C_{11})^{-1} K^0 \tau^0$$

Multiplying both sides of the equation by $(K^0 \tau^0)^T = (\tau^0)^T K^0$ we obtain

$$(D.6) \quad 1 = -(\tau^0)^T K^0 (C_{11})^{-1} C_{12} \Delta_{(2)}^0 + \lambda (K^0 \tau^0)^T (C_{11})^{-1} (K^0 \tau^0)$$

Note that the absolute value of the first term,

$$(D.7) \quad \left| \sum_{g \notin S} \left(\Delta_{[g]}^0\right)^T \left[C_{21} (C_{11})^{-1} K^0 \tau^0 \right]_{[g]} \right|,$$

is bounded above by

$$(D.8) \quad (1 - \eta) \left(\sum_{g \notin S} \sqrt{k_g} \|\Delta_{[g]}^0\|_2 \right) \leq (1 - \eta) L$$

by virtue of the uniform irrepresentable condition and the Cauchy-Schwartz inequality.

Assuming the minimum eigenvalue of C_{11} , i.e., $\Lambda_{\min}(C_{11})$, is positive and considering $\|K^0 \tau^0\|_2 \leq \sqrt{q}$, the second term is at most $\lambda q / \Lambda_{\min}(C_{11})$. So (D.6) implies

$$(D.9) \quad 1 \leq (1 - \eta) L + \frac{\lambda q}{\Lambda_{\min}(C_{11})}$$

In particular, $\lambda \geq \Lambda_{\min}(C_{11}) (1 - (1 - \eta) L) / q$ is positive whenever $L < 1/(1 - \eta)$.

Next, multiply both sides of (D.5) by $(\Delta_{(2)}^0)^T C_{21}$ to get

$$(D.10) \quad \left(\Delta_{(2)}^0\right)^T C_{21} \Delta_{(1)}^0 + \left(\Delta_{(2)}^0\right)^T C_{21} (C_{11})^{-1} C_{12} \Delta_{(2)}^0 = \lambda \left(\Delta_{(2)}^0\right)^T C_{21} (C_{11})^{-1} K^0 \tau^0$$

Using the upper bound in (D.8), the right hand side is at least $-\lambda(1-\eta)L$. Also a simple consequence of the block inversion formula of the non-negative definite matrix C guarantees that the matrix $C_{22} - C_{21}(C_{11})^{-1}C_{12}$ is non-negative definite. Hence,

$$\begin{aligned} & \left(\Delta_{(2)}^0\right)^T \left[C_{22} - C_{21}(C_{11})^{-1}C_{12} \right] \Delta_{(2)}^0 \geq 0 \\ \text{and } & \left(\Delta_{(2)}^0\right)^T C_{22}\Delta_{(2)}^0 \geq \left(\Delta_{(2)}^0\right)^T C_{21}(C_{11})^{-1}C_{12}\Delta_{(2)}^0 \end{aligned}$$

Putting all the pieces together we get

$$\begin{aligned} \phi_{compatible}^2/q &= \frac{1}{n} \|\mathbf{X}\Delta^0\|_2^2 \\ &= \Delta_{(1)}^0{}^T C_{11}\Delta_{(1)}^0 + 2\Delta_{(2)}^0{}^T C_{21}\Delta_{(1)}^0 + \Delta_{(2)}^0{}^T C_{22}\Delta_{(2)}^0 \\ &= \lambda + \Delta_{(2)}^0{}^T C_{21}\Delta_{(1)}^0 + \Delta_{(2)}^0{}^T C_{22}\Delta_{(2)}^0, \text{ by (D.4)} \\ &\geq \lambda - \lambda(1-\eta)L, \text{ by (D.10)} \\ &= \lambda(1 - (1-\eta)L) \end{aligned}$$

Plugging in the lower bound for λ we obtain the result; namely,

$$\phi_{compatible}^2 = \Lambda_{min}(C_{11}) (1 - (1-\eta)L)^2 > 0$$

for any $L < \frac{1}{1-\eta}$. □

D.1. Necessity of the Weak Irrepresentable Condition for direction consistency. In this section we demonstrate the necessity of weak irrepresentable condition for group sparsity selection and direction consistency. We shall assume that the minimum eigenvalue of the signal part of the Gram matrix, viz. $\Lambda_{min}(C_{11})$, is bounded below. We shall also assume that the matrices C_{21} and C_{22} are bounded above in spectral norm. Suppose that the weak irrepresentable condition does not hold, i.e., for some $g \notin S$ and $\xi > 0$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1}K^0\tilde{D}(\beta_{(1)}^0) \right]_{[g]} \right\| > 1 + \xi$$

for infinitely many n . Also suppose that there exists a sequence of positive reals $\delta_n \rightarrow 0$ such that the event

$$E_n := \{ \|D(\hat{\beta}_{[g]}) - D(\beta_{[g]})\|_2 < \delta_n, \forall g \in S, \text{ and } \hat{\beta}_{[g]} = \mathbf{0} \forall g \notin S \}$$

satisfies $\mathbb{P}(E_n) \rightarrow 1$ as $p, n \rightarrow \infty$.

Note that for large enough n so that $\delta_n < \min_g \|D(\beta_{[g]})\|$, we have $\hat{\beta}_{[g]} \neq \mathbf{0}$, $\forall g \in S$ on the event E_n .

Then, as in the proof of Theorem 4.6, we have, on the event E_n ,

$$(D.11) \quad \hat{\mathbf{u}} = (C_{11})^{-1} \left[\frac{1}{\sqrt{n}} Z_{(1)} - \lambda K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]$$

$$(D.12) \quad \text{and} \quad \frac{1}{n} \left\| [\mathbf{X}_{(2)}^T (\epsilon - \mathbf{X}_{(1)} \hat{\mathbf{u}})]_{[g]} \right\| \leq \lambda \sqrt{k_g}, \quad \forall g \notin S$$

Substituting the value of $\hat{\mathbf{u}}$ from (D.11) in (D.12), we have, on the event E_n ,

$$\frac{1}{\sqrt{n}} \left\| \left[\mathbf{Z}_{(2)} - C_{21}(C_{11})^{-1} \mathbf{Z}_{(1)} + \lambda \sqrt{n} C_{21}(C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| \leq \lambda \sqrt{k_g},$$

which implies that

$$(D.13) \quad \begin{aligned} & \left\| \left[\mathbf{Z}_{(2)} - C_{21}(C_{11})^{-1} \mathbf{Z}_{(1)} \right]_{[g]} \right\| \\ & \geq \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| - 1 \right]. \end{aligned}$$

Now note that for large enough n , if $\|C_{21}\|$ is bounded above, direction consistency guarantees that the expression on the right is larger than

$$\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\beta_{(1)}) \right]_{[g]} \right\| - 1 \right]$$

which in turn is larger than $\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \xi$, in view of the weak irrerepresentable condition.

This contradicts $\mathbb{P}(E_n) \rightarrow 1$, since the left-hand side of (D.13) corresponds to the norm of a zero mean Gaussian random variable with bounded variance structure $[C_{22} - C_{21}(C_{11})^{-1} C_{12}]_{[g][g]}$ while the right hand side diverges for

$$\lambda \asymp \frac{\sqrt{s}(\sqrt{k_g} + \sqrt{\log G})}{\sqrt{n}}.$$

REFERENCES

- BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. [MR2417268 \(2010a:68132\)](#)
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732.
- BLANCHARD, O. and PEROTTI, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *the Quarterly Journal of economics* **117** 1329–1368.

- BREHENY, P. and HUANG, J. (2009). Penalized methods for bi-level variable selection. *Stat. Interface* **2** 369–380. [MR2540094 \(2010k:62290\)](#)
- FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science's STKE* **303** 799.
- FUJITA, A., SATO, J., GARAY-MALPARTIDA, H., YAMAGUCHI, R., MIYANO, S., SOGAYAR, M. and FERREIRA, C. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology* **1** 39.
- GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix computations*, third ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, MD. [MR1417720 \(97g:65006\)](#)
- GRANGER, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37** 424–438.
- HIEMSTRA, C. and JONES, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance* 1639–1664.
- HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978–2004. [MR2676881](#)
- HUANG, J., MA, S., XIE, H. and ZHANG, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. [MR2507147](#)
- KIM, K., KIM, J. H., LEE, J., JIN, H. M., LEE, S. H., FISHER, D. E., KOOK, H., KIM, K. K., CHOI, Y. and KIM, N. (2005). Nuclear factor of activated T cells c1 induces osteoclast-associated receptor gene expression during tumor necrosis factor-related activation-induced cytokine-mediated osteoclastogenesis. *Journal of Biological Chemistry* **280** 35209–35216.
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach spaces. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer-Verlag, Berlin. Isoperimetry and processes. [MR1102015 \(93c:60001\)](#)
- LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204.
- LOZANO, A., ABE, N., LIU, Y. and ROSSET, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics* **25** i110.
- LÜTKEPOHL, H. (2005). *New introduction to multiple time series analysis*. Springer.
- NARDI, Y. and RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605–633. [MR2426104 \(2009k:62175\)](#)
- PARTER, S. V. (1961). Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations. *Trans. Amer. Math. Soc.* **99** 153–192. [MR0120492 \(22 ##11245\)](#)
- PEARL, J. (2000). *Causality: models, reasoning, and inference* **47**. Cambridge Univ Press.
- PRIESTLEY, M. B. (1981). *Spectral analysis and time series. Vol. 2*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London. Multivariate series, prediction and control, Probability and Mathematical Statistics. [MR628736 \(83b:62186b\)](#)
- RANGEL, C., ANGUS, J., GHAHRAMANI, Z., LIOUMI, M., SOTHERAN, E., GAIBA, A., WILD, D. L. and FALCIANI, F. (2004). Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics* **20** 1361.
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. [MR2719855 \(2011h:62272\)](#)
- RUDELSON, M. and ZHOU, S. (2011). Reconstruction from anisotropic random measurements. *Arxiv preprint arXiv:1106.1151v1*.
- SHOJAIE, A., BASU, S. and MICHAELIDIS, G. (2012). Adaptive Thresholding for Reconstructing Regulatory Networks from Time-Course Gene Expression Data. *Statistics in Biosciences* **4** 66–83. 10.1007/s12561-011-9050-5.

- SHOJAIE, A. and MICHAILIDIS, G. (2010a). Penalized Likelihood Methods for Estimation of Sparse High Dimensional Directed Acyclic Graphs. *Biometrika* **97** 519–538.
- SHOJAIE, A. and MICHAILIDIS, G. (2010b). Discovering Graphical Granger Causality Using a Truncating Lasso Penalty. *Bioinformatics* **26** i517–i523.
- SIMS, C. A. (1972). Money, income, and causality. *The American Economic Review* **62** 540–552.
- VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392.
- VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–749. . [MR2820636](#)
- WATERMAN, M., JONES, K. et al. (1990). Purification of TCF-1 alpha, a T-cell-specific transcription factor that activates the T-cell receptor C alpha gene enhancer in a context-dependent manner. *The New biologist* **2** 621.
- WEI, F. and HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369–1384. . [MR2759183](#)
- ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. . [MR2549566 \(2011c:62234\)](#)
- ZHAO, P. and YU, B. (2006). On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563.
- ZHOU, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. *Arxiv preprint arXiv:1002.1583*.

SUMANTA BASU AND GEORGE MICHAILIDIS
 DEPARTMENT OF STATISTICS
 UNIVERSITY OF MICHIGAN
 ANN ARBOR MI 48109
 E-MAIL: sumbose@umich.edu
gmichail@umich.edu

ALI SHOJAIE
 DEPARTMENT OF BIostatISTICS
 UNIVERSITY OF WASHINGTON
 SEATTLE WA 98195
 E-MAIL: ashojaie@u.washington.edu

TABLE 1

Performance of different regularization methods in estimating graphical Granger causality with **balanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60, E = 351$ Group Size=3			$p = 120, E = 1404$ Group Size=3			$p = 200, E = 3900$ Group Size=5		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	80(2)	75(2)	66(4)	69(1)	62(2)	52(2)	52(1)	47(1)	38(1)
	Grp	95(2)	91(4)	83(7)	91(3)	80(5)	68(7)	78(4)	72(3)	59(6)
	Thgrp	96(1)	92(3)	86(6)	93(3)	83(5)	70(7)	82(4)	76(3)	64(6)
	Agrp	96(2)	92(4)	83(7)	92(3)	82(5)	69(7)	81(3)	74(3)	60(6)
R	Lasso	71(2)	54(2)	31(2)	54(1)	40(1)	22(1)	38(1)	28(1)	15(1)
	Grp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(1)	70(2)	41(4)
	Thgrp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(2)	69(2)	41(3)
	Agrp	99(1)	93(3)	71(7)	91(2)	81(2)	47(8)	84(1)	69(2)	40(4)
MCC	Lasso	75(2)	63(2)	45(3)	60(1)	49(1)	33(1)	43(1)	35(1)	23(1)
	Grp	97(1)	92(3)	76(5)	91(1)	80(2)	56(2)	81(2)	70(2)	48(2)
	Thgrp	98(1)	93(2)	78(5)	92(1)	81(2)	57(3)	83(2)	72(2)	50(3)
	Agrp	97(1)	92(3)	76(5)	91(1)	81(2)	56(3)	82(2)	71(2)	48(2)
ERR LAG	Lasso	10.5	11.3	13.9	16.63	17.37	16.69	19.79	20	18.52
	Grp	3.19	6.95	12.76	4.86	10.77	12.65	4.21	5.27	7.8
	Thgrp	2.83	5.87	10.01	3.98	9.03	11.19	3.06	3.91	5.68
	Agrp	3.13	6.89	12.59	4.63	10.37	12.34	3.58	4.87	7.59

TABLE 2

Performance of different regularization methods in estimating graphical Granger causality with **unbalanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60, E = 450$ Groups=1 \times 10, 11 \times 5			$p = 120, E = 1575$ Groups=1 \times 10, 23 \times 5			$p = 200, E = 4150$ Groups=1 \times 10, 39 \times 5		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	72(2)	69(3)	62(2)	51(1)	48(1)	41(1)	61(1)	53(1)	42(2)
	Grp	84(4)	79(6)	76(9)	55(5)	47(5)	40(6)	86(3)	77(5)	66(7)
	Thgrp	86(4)	82(7)	78(11)	60(6)	50(7)	40(5)	88(2)	79(6)	69(6)
	Agrp	85(3)	81(5)	77(9)	59(5)	51(5)	42(6)	88(2)	78(5)	67(6)
R	Lasso	45(2)	35(2)	22(2)	43(1)	34(1)	22(1)	23(1)	15(0)	7(0)
	Grp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
	Thgrp	95(2)	88(4)	62(8)	89(3)	77(4)	50(5)	73(3)	50(6)	21(5)
	Agrp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
MCC	Lasso	56(2)	48(2)	35(2)	46(1)	39(1)	29(1)	36(1)	28(1)	17(1)
	Grp	89(3)	82(4)	67(5)	68(3)	58(3)	42(3)	79(1)	61(3)	37(3)
	Thgrp	90(3)	84(4)	68(6)	72(4)	61(4)	43(2)	80(1)	62(3)	37(3)
	Agrp	89(3)	83(4)	67(6)	71(3)	60(3)	43(3)	79(1)	61(3)	37(3)
ERR LAG	Lasso	10.59	10.74	11.76	18.3	18.72	18.76	11.54	10.93	9.29
	Grp	7.04	9.85	13.04	12.53	14.71	13.06	4.8	6.41	6.85
	Thgrp	6.58	8.98	11.1	9.6	11.9	10.9	4.06	5.65	5.7
	Agrp	6.74	9.19	12.96	10.81	12.78	11.79	4.55	6.2	6.81

TABLE 3

Performance of different regularization methods in estimating graphical Granger causality with **misspecified** groups (30% misspecification); $d = 2$, $T = 10$, $SNR = 2$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60, E = 246$			$p = 120, E = 968$			
		Group Size=6			Group Size=6			
		n	160	110	60	160	110	60
P	Lasso	88(2)	85(3)	77(5)	59(1)	55(1)	49(2)	
	Grp	65(2)	66(2)	66(3)	43(3)	44(4)	38(4)	
	Thgrp	87(3)	88(3)	85(3)	56(6)	56(6)	51(7)	
	Agrp	65(2)	66(2)	66(3)	45(2)	45(4)	39(4)	
R	Lasso	80(3)	63(3)	37(2)	66(1)	54(1)	35(1)	
	Grp	100(0)	98(2)	82(6)	87(2)	78(3)	59(4)	
	Thgrp	100(0)	98(2)	79(6)	86(2)	79(3)	57(4)	
	Agrp	100(0)	98(2)	82(6)	86(2)	78(3)	58(3)	
MCC	Lasso	84(2)	73(2)	53(3)	62(1)	54(1)	41(1)	
	Grp	81(1)	80(2)	74(4)	61(2)	58(3)	47(2)	
	Thgrp	93(2)	93(2)	82(4)	69(4)	66(4)	53(3)	
	Agrp	81(1)	80(2)	74(4)	62(2)	59(2)	47(2)	
ERR LAG	Lasso	12.63	17.05	22.41	45.09	49.68	53.4	
	Grp	9.43	8.78	15.12	18.22	18.43	29.26	
	Thgrp	6.45	5.34	8.02	11.81	12.84	15.57	
	Agrp	9.11	8.78	14.96	16.32	16.9	27.69	

TABLE 4

Mean and standard deviation (in parentheses) of PMSE (MSE in case of Dec 2010) for prediction of banking balance sheet variables.

Quarter	Lasso	Grp	Agrp	Thgrp
Dec 2010	1.59 (0.29)	0.36 (0.05)	0.36 (0.05)	0.37 (0.05)
Mar 2011	1.46 (0.30)	0.47 (0.23)	0.47 (0.23)	0.46 (0.22)
Jun 2011	1.33 (0.26)	0.36 (0.11)	0.36 (0.11)	0.35 (0.11)
Sep 2011	1.72 (0.32)	0.50 (0.18)	0.50 (0.18)	0.47 (0.16)

TABLE 5

Mean and standard deviation of MSE for different NGC estimates

	Lasso	Grp	Agrp	Thgrp
mean	0.649	0.456	0.457	0.456
stdev	0.340	0.252	0.251	0.252

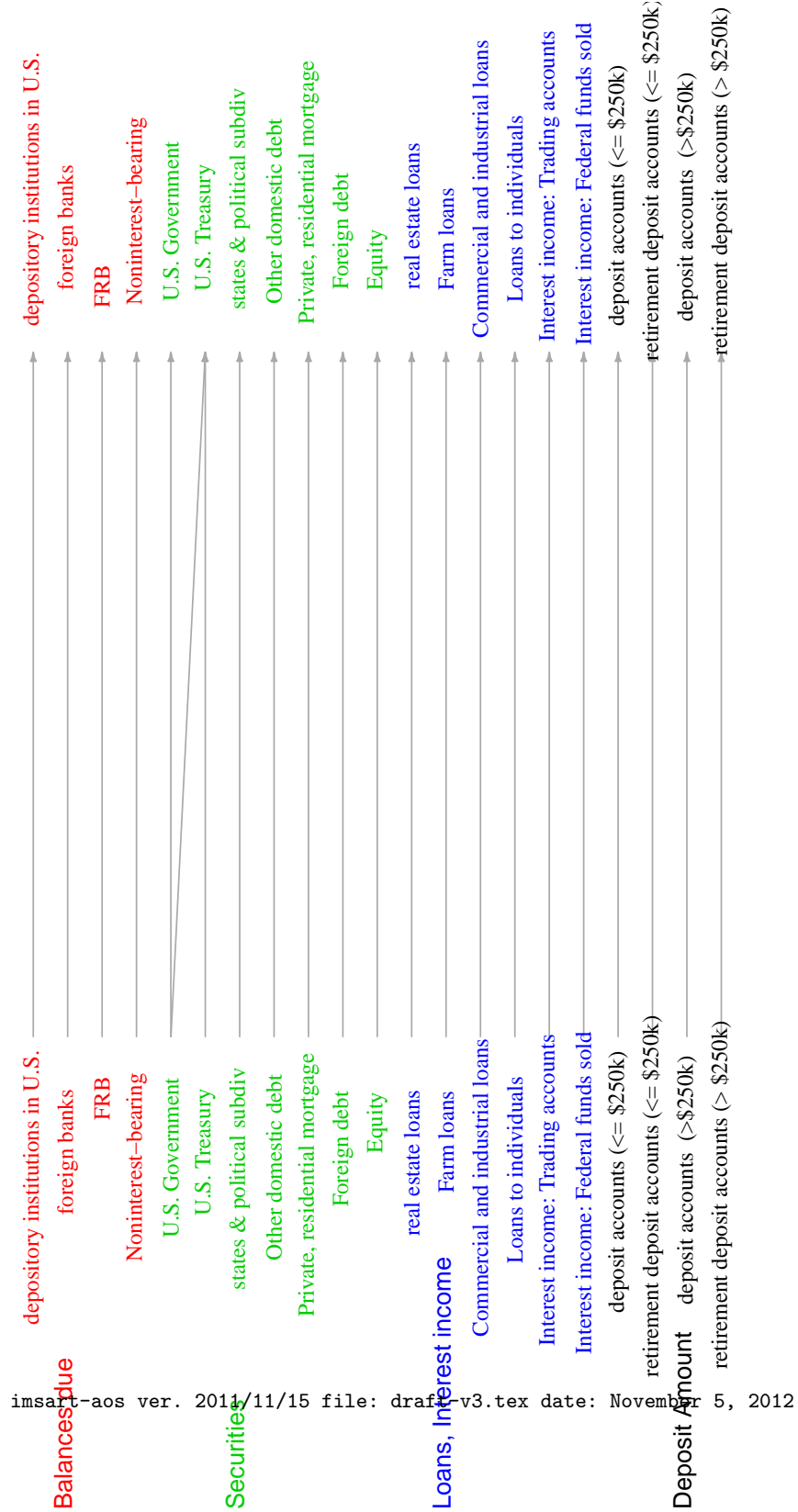


Fig 3: Estimated Networks of banking balance sheet variables using lasso. The network represents the aggregated network over 5 time points.

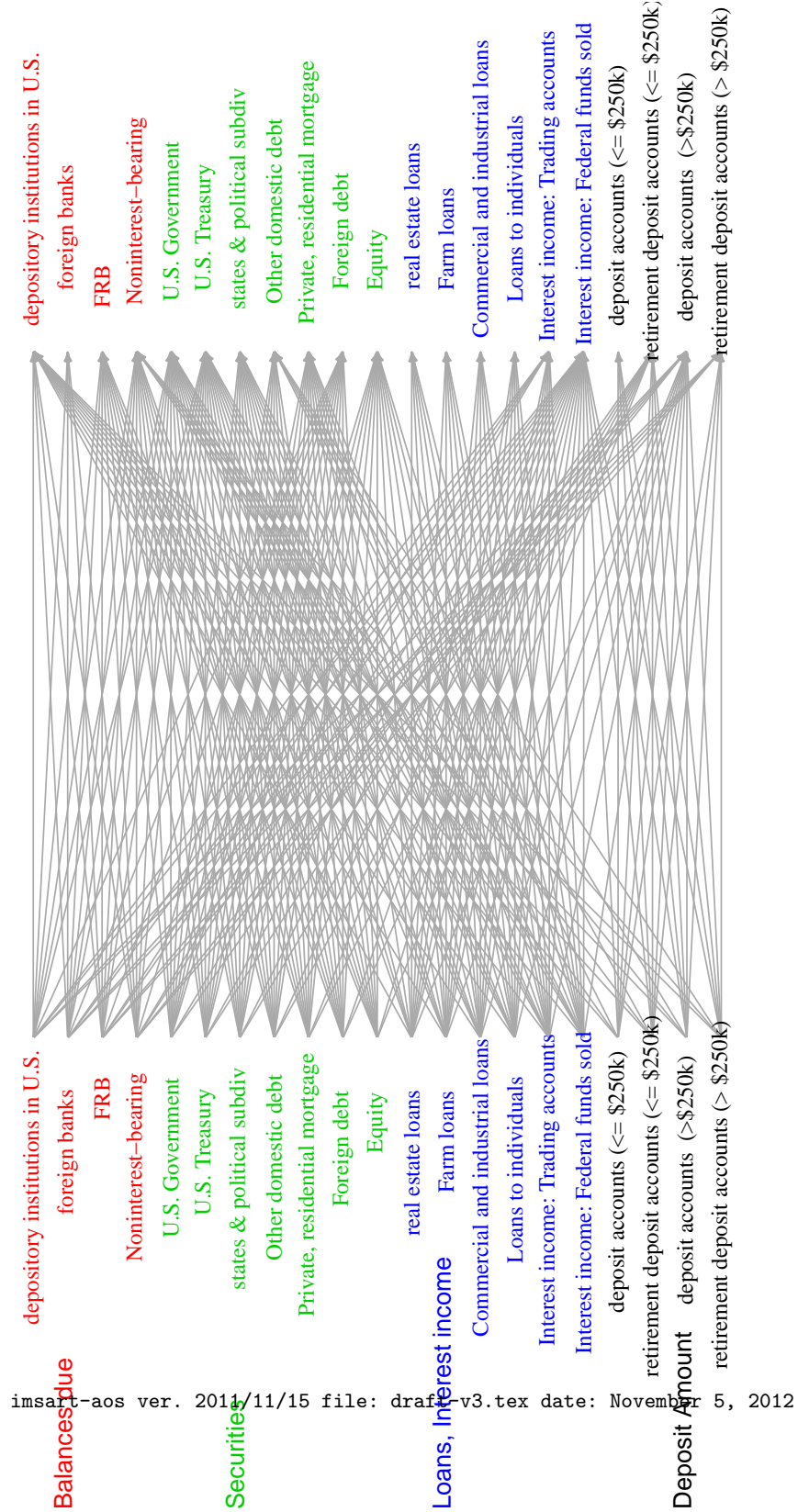


Fig 4: Estimated Networks of banking balance sheet variables using group lasso. The network represents the aggregated network over 5 time points.

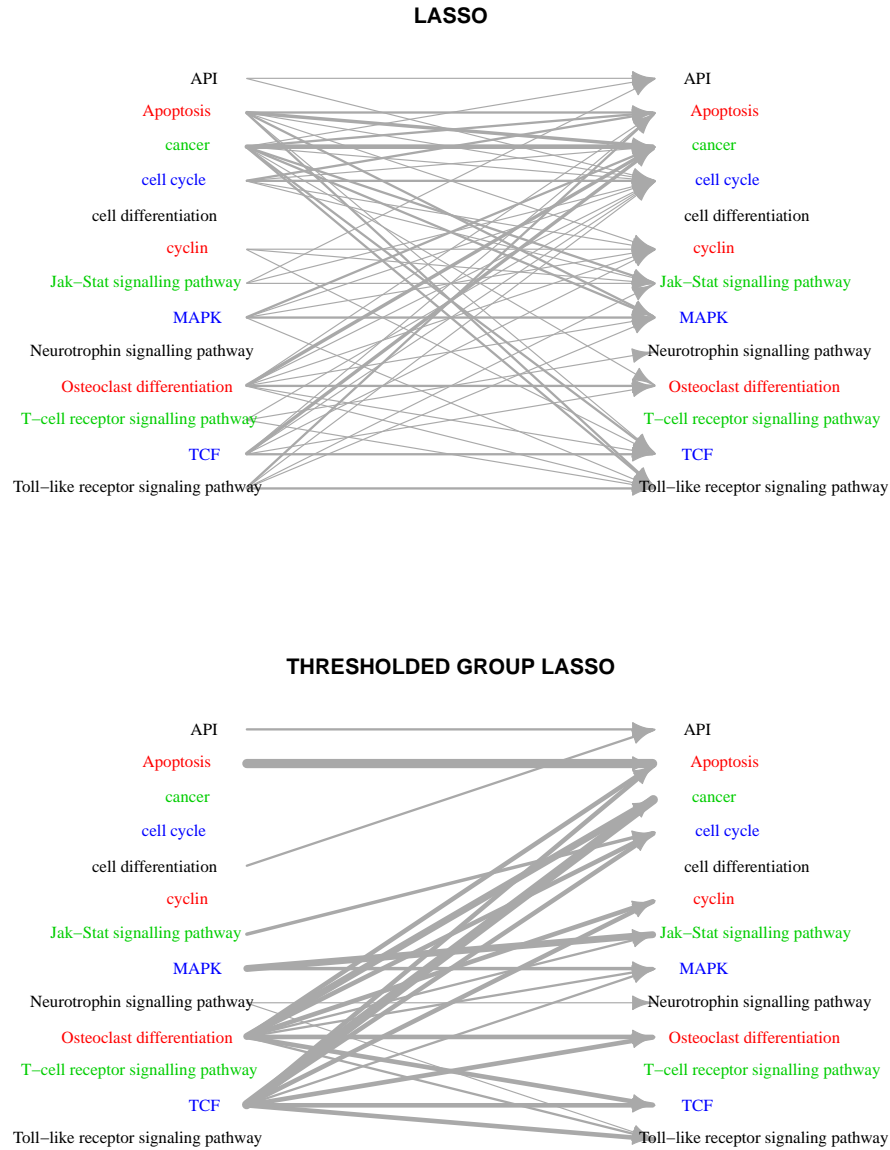


Fig 5: Estimated Gene Regulatory Networks of T-cell activation. Width of edges represent the number of effects between two groups, and the network represent the aggregated regulatory network over 3 time points.